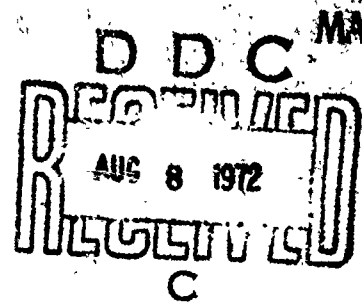




FRANK J. SEILER RESEARCH LABORATORY

SRL-TR-72-0004

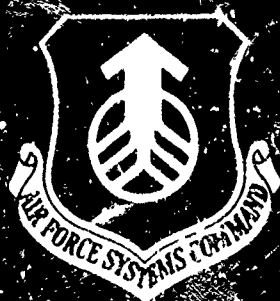
AD 746921



AN ENGINEER'S GUIDE  
TO BUILDING NONLINEAR FILTERS

VOLUME I

Richard S Bucy  
Calvin Hecht  
Capt Kenneth D Sonne



PROJECT 7904

APPROVED FOR PUBLIC RELEASE;  
DISTRIBUTION UNLIMITED

AIR FORCE SYSTEMS COMMAND  
UNITED STATES AIR FORCE

302

(UNCLASSIFIED)

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)		2a. REPORT SECURITY CLASSIFICATION	
Frank J. Seiler Research Laboratory (AFSC) USAF Academy, Colorado 80840		UNCLASSIFIED	
3. REPORT TITLE		2b. GROUP	
An Engineer's Guide to Building Nonlinear Filters			
4. DESCRIPTIVE NOTES (Type of report and inclusive dates)			
Final Project Report (1969-1972)			
5. AUTHOR(S) (First name, middle initial, last name)			
Richard S. Bucy, Professor, U.S.C. Calvin Hecht, T.R.W. Systems Kenneth D. Senne, Captain, USAF			
6. REPORT DATE		7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
May 1972		541	44
8a. CONTRACT OR GRANT NO.		9a. ORIGINATOR'S REPORT NUMBER(S)	
b. PROJECT NO. 7904-00-17		SRL-TR-72-0004	
c. DRS 61102F		9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	
d. BPAC 681307		AD-	
10. DISTRIBUTION STATEMENT			
Approved for public release; distribution unlimited.			
11. SUPPLEMENTARY NOTES		12. SPONSORING MILITARY ACTIVITY	
		Frank J. Seiler Research Laboratory (AFSC) USAF Academy, Colorado 80840	
13. ABSTRACT			
<p>A comprehensive treatment of numerical approaches to the solution of Bayes Law has been included, describing numerical methods, computational algorithms, two example problems, and extensive numerical results. Bayes Law is an integral equation describing the evolution of the conditional probability distribution, describing the state of a Markov process, conditioned on the past noisy observations. The Bayes Law is, in fact, the general solution to the discrete nonlinear estimation problem. This research represents one of the first successful attempts to approximate the conditional probability densities numerically and evaluate the Bayes integral by quadratures. The methods of density representation studied most thoroughly include Orthogonal Polynomials, Point-Masses, Gaussian Sums, and Fourier Series.</p> <p>For example problems two second-order systems have been studied. The first problem involves a passive (bearings-only) receiver with geometry similar to the AWACS. The second example involves the reconstruction of a second order phase-process which is the message process for a phase-modulated communication system. The various forms of the nonlinear estimates are compared with the phase-locked loop demodulator and extensive Monte Carlo simulations are described to provide high confidence numerical comparisons.</p> <p>A chapter is devoted to elaborating on the Monte Carlo methods employed for the computer simulations and a general-purpose, high-quality random number generator is introduced which is exactly realizable on any binary computer for comparisons of the experimental results. Another chapter discusses the applicability of parallel digital and hybrid computer architectures to the Bayes-Law algorithms.</p>			

DD FORM 1473  
1 NOV 65

1a

UNCLASSIFIED  
Security Classification

UNCLASSIFIED  
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Markov Processes						
Nonlinear estimation						
Filtering						
Prediction						
Time series						
Simulation						
Bayes Law						
Monte Carlo methods						
Orthogonal Series						
Quadratures						
Numerical Curve Fitting						
Random Number Generators						
Parallel Processing						
Hermite Polynomials						
Gaussian Sums						
Point Masses						
Passive Tracking						
Bearings-only Receivers						
AWACS						
Phase Modulation						
Demodulation						
Phase-Locked Loops						
<i>if</i>						

An Engineer's Guide  
to Building  
Nonlinear Filters

by

Richard S. Bucy, U.S.C.

Calvin Hecht, T.R.W. Systems

Kenneth D. Senne, Capt., U.S.A.F.\*

Final Report  
Frank J. Seiler Research Laboratory  
Project 7904-00-37  
May 1972

---

\* After June 1972: Staff Member, M.I.T. Lincoln Laboratory

Publication Notice

The authors are submitting portions of the enclosed material to the IEEE Transactions on Automatic Control, while other portions will be published in Stochastics.

## Foreword

Ever since the publication of the papers by Kalman and Bucy in 1960 and 1961, the engineering community has made giant strides in applying their ideas concerning linear estimation to an essentially uncountable number of space and control problems. The widespread and almost instantaneous acceptance of the Kalman-Bucy filter was due in part to a number of factors: the then common frequency domain synthesis procedures were limited to time-invariant (steady-state) problems, the advent of the digital computer led to easy simulation of time functions, whereas frequency calculations were indirect and less convenient than "state-space", and most important, the newly accelerated man-in-space program led to an immediate real-time application, the most significant catalyst for engineering developments.

The fact that most applications were not linear systems did not diminish the enthusiasm for the new linear theory. Almost over night, under the pressures of ever-present engineering deadlines, a host of approximate approaches to the nonlinear problems were generated with many variations and extensions, but most of which were patterned to look much like the highly successful linear estimator of Kalman and Bucy. In some applications these approximations produced highly satisfactory results, but in most situations a second design phase, less publicized but equally important, was carried out in order to eliminate the anomalies, instabilities and unexplained peculiarities of the estimators. A seemingly endless stream of technical papers dealing with adaptive techniques and examples of clever but non-generalizable

"engineering fixes" for the multitude of undesirable characteristics of the approximations have appeared during the last 8-10 years, resulting in an apparently bottomless bag of engineering tricks for the estimation engineer. In addition, the technical access to the original articles was evidently unsatisfactory since, for a variety of reasons, many authors have devoted tutorial articles explaining how the same equations for estimation can be derived from a variety of points of view, including Bayes rule (the original concept), least squares, maximum likelihood and a host of others. The results would fill many volumes of "handbooks" for the engineer while neither answering nor posing the questions - where do we go from here? or - do we really understand nonlinear estimation?

On the other hand, it is fair to say that nonlinear estimation theory, as advanced and generalized as it has become, has not been publicized and advertised as a general panacea for the estimation problems of the future. A small but growing contingent of university researchers has been working steadily on problems involving mathematically subtle concepts such as stochastic integrals, Ito calculus, diffusion processes, etc., and a few elegant and mathematically sophisticated theorems have been proved concerning general representations for nonlinear estimators. But the very difficulties which frustrated the application of these "solutions" is their nature: estimates are determined to be related to the numerical solution of infinite-dimensional partial differential equations - a formidable task in the simplest of problems, and thus, understandably, the paths of estimation application and nonlinear estimation theory have continued to diverge. On the rare occasion that representatives of the two

factions have met much needless antagonism has resulted. While it is true that the questions most frequently asked by the applications engineers rarely deal with the optimality of estimates but merely with the computational tractability, and vice versa for the theorists, it appears that a partial reconciliation on both sides seems likely to provide mutual enrichment.

Specifically, what is proposed here is that all interested parties should stand back for a moment and reflect on the question - what, exactly, are the characteristics of optimal, nonlinear estimators? The answer to this and related questions provides the motivation for the work described in this paper. We ask that the applications engineer temporarily suspend his seemingly unattainable computational constraints and that the estimation theorist pause briefly from his esoteric pursuits involving the subtleties of measure theory and look for awhile at some examples of optimal, discrete-time, nonlinear estimators. (As it turns out, the discrete-time problem, although far from trivial, is at least tractable for solution on modern digital computers.) In this way, it is expected that the benefits to both factions will be considerable, and perhaps their steady divergence may be curtailed.

In effect, it is hoped that the engineer engaged in the daily process of fitting old tricks into new computers will begin to ask whether or not the basic principle of the trick is applicable to the specific application at hand - perhaps a simple modification or another approach would be much more satisfactory. Such realizations are generally obtained at the expense of serious time delays and costly experimental failures - perhaps added experience with optimal nonlinear estimators could provide more effective and less expensive guidelines.

In addition, it is hoped that the theoretician can appreciate the reality of concrete examples and thereby refuel his fire regarding theorems concerning the characteristics and asymptotic descriptions of optimal nonlinear estimates. Finally, it is intended that certain applications of nonlinear estimators be considered for which the optimal solution can itself be considered practical, if not via present day realizations, then perhaps by special purpose hardware designed with optimal estimation in mind. This paper has been written to instill some enthusiasm in the reader for all of these expectations.

### Acknowledgements

The authors are deeply indebted to the generous and enthusiastic support of the U.S. Air Force and to the National Aeronautics and Space Administration. The research described herein has been performed at the Frank J. Seiler Research Laboratory, Air Force Systems Command, at the University of Southern California Department of Aerospace Engineering (Air Force Office of Scientific Research Grant AFOSR-71-2141), and in conjunction with Electrac Corporation (NASA Grant NAS5-10789, 1970). Although many persons must be acknowledged regarding this research and the associated experiments, perhaps the two most important figures are Colonel Bernard S. Morgan, Jr., USAF, and Major Allen D. Dayton, USAF, who had the insight and intuition necessary to introduce the authors of this paper to each other in the first place. It would be very fortunate, indeed, if such technical interchanges and mutual cooperation were to continue to be encouraged among government sponsored and university researchers.

## Table of Contents

	<u>Page</u>
Publication Notice	ii
Foreward	iii
Acknowledgements	vii
Table of Contents	ix
Table of Figures	xiii
Table of Tables	xvi
I. Introduction	1
References	6
ii. Bayesian Estimation: The Problem	9
References	21
III. Finite Dimensional Approximations	23
A. Introduction	23
B. Orthogonal Series	24
1. Least Square Polynomial Approximation, Scalar Case	25
2. Least Square Polynomial Approximation, Multi-dimensional Case	31
3. Gauss-Hermite Integration	34
4. Application of Polynomial Expansions to a Two- Dimensional Filtering Problem	39
5. Applying the Hermite Expansion	50

*Preceding page blank*

	<u>Page</u>
C. The Point-Mass Approximation	64
D. Non-Orthogonal Series - Gaussian Sums	68
E. Other Computational Methods	74
1. Fourier Series Expansions	74
2. Spline Functions	75
References	77
IV. Monte Carlo Analysis Techniques	79
A. Introduction	79
B. Statistical Analysis	80
C. An Example	85
D. Conclusions	94
References	96
Appendix. A Machine Independent Random Number Generator	97
A. Introduction	97
B. An Example	100
V. Parallel Computational Techniques	115
A. Introduction	115
B. Parallelism and Bayes Law	115
C. Look-Ahead Processors	122
D. Array Processors	124
E. Associative Processors	125
F. Pipe-Line Processor	125
G. Hybrid Computer Methods	127
References	129
VI. A Passive Receiver: Bearings-Only Tracking (AWACS)	131
A. Introduction	131

	<u>Page</u>
B. The Linearized Estimator	136
C. Application of Nonlinear Filtering	137
References	140
Appendix A. First Monte Carlo Experiments - Point Masses Versus Linearized	141
Appendix B. More Recent Experimental Results: Point Masses Versus Gaussian Sums	147
Appendix C. A Movie of Conditional Densities	155
VII. Example: Optimal nonlinear Phase Demodulation	183
A. Introduction	183
B. The Linearized Filter	184
C. Application of Nonlinear Filtering	198
Referenced	205
Appendix A. Numerical Experiments with the Phase- Locked Loop	207
Appendix B. Numerical Experiments with The Hermite Polynomial Expansion	213
Appendix C. Cyclic Point-Mass Experiments	227
Appendix D. A Fourier Series Experiment	243
Appendix E. A Movie of Conditional Densities	249
VIII. Conclusions	267
Bibliography	271
Resumes of the Authors	275
Richard S. Bucy	276
Calvin Hacht	281
Kenneth D. Senne	282

	<u>Page</u>
Additional Appendix A. A Two-Dimensional Point-Mass Program for the Passive Receiver Problem	285
Additional Appendix B. A Gaussian-Sum Program for the Passive Receiver Problem	355
Additional Appendix C. A Gauss-Hermite Program for Implementing the Two-Dimensional Phase Demodulator	379
Additional Appendix D. A Point-Mass Program for Implementing the Interpolating Version of the Cyclic Phase Demodulator	419
Additional Appendix E. A Fourier Series Implementation of the Cyclic Phase Demodulator	445
Additional Appendix F. Some Unpublished Conference Papers Referenced by this Report	465
R.S. Bucy, "Realization of Non-Linear Filters"	467
R.S. Bucy, M.J. Merritt, and D.S. Miller, "Hybrid Computer Synthesis of Optimal Discrete Nonlinear Filters"	475
C. Hecht, "Digital Realization of Non-Linear Filters"	505
K.D. Senne, "Computer Experiments with Nonlinear Filters"	513

## Table of Figures

	<u>Page</u>
 Chapter III	
Fig. 1. Hermite Polynomial Bayes-Law Recursion	41
Fig. 2. Coordinate Systems for Hermite Expansion	58
 Chapter IV	
Fig. 1. Example of a Questionable Monte Carlo Cumulative Average Sample Path	84
Fig. 2. Probability Density and Distribution of the Asymptotic ( $N \rightarrow \infty$ ) Kolmogorov Statistic	91
Fig. A-1. Algorithm for Partitioned Uniform Generator	101
 Chapter V	
Fig. 1. Serial Evaluation of $a+b+c+d+e+f$	117
Fig. 2. Maximally Parallel Evaluation of $a+b+c+d+e+f$	118
Fig. 3. Combining Serial and Parallel Computations	121
 Chapter VI	
Fig. 1. Typical Passive Receiver Geometry	133
Fig. B-1. Typical Geometry of the "Old Problem" - Illustrates Periodicity of Errors	149
Fig. B-2. "New Problem" Geometry without Periodic Errors	151
Fig. C-1. Detection Geometry in the Presence of Multipath Reception	156
Fig. C-2. A Priori Density Resulting from Multipath Detection Ambiguity	158

	<u>Page</u>
Fig. C-3. A Typical Sample Path Resulting from the Multipath Detection Ambiguity	159
Fig. C-4. Absolute Error Performance of Optimal and Linearized Predictors for Multimodal Problem	181
 Chapter VII	
Fig. 1. Block Diagram of Linearized Phase Estimation	189
Fig. 2. Discrete $P_{11}$ Error Variance	195
Fig. 3. Discrete $P_{22}$ Error Variance	196
Fig. 4. Discrete $P_{12}$ Error Variance	197
Fig. 5. Torus Interpretation of Doubly Cyclic State Space	201
Fig. A-1. MSE Performance Summary	209
Fig. A-2. Fourth Moment Divided by Three Times the Squared Variance for the Phase-Locked Loop Error	211
Fig. B-1. Hermite Expansion Error Summary - $P(o) = 4P(\infty)$	218
Fig. B-2. Cumulative Statistical Variance - $P(o) = 4P(\infty)$	219
Fig. B-3. Portion of Sample Function No. 6 - $P_{11}(o) = 0.3025$	220
Fig. B-4. Error for Sample Function No. 6 - $P_{11}(o) = 0.3025$	221
Fig. B-5. Error Variance for $P_{11}(o) = -5.2$ dB Starting at $P_{11}(o) = 4P_{11}(\infty)$	223
Fig. B-6. Cumulative Statistical Variance for $P_{11}(o) = -5.2$ dB	224
Fig. C-1. Nonlinear Filter Summary (Enlarged)	231
Fig. C-2. MSE Improvement of Nonlinear Filters over Phase-Locked Loop	235

	<u>Page</u>
Fig. C-3. MSE Difference from Ideal Linear Analysis	237
Fig. E-1. A Typical Sample Path of Densities Evolving in Time	250

## Table of Tables

	<u>Page</u>
 Chapter II	
Table 1. Model for Bayes Rule Conditional Density	
Recursion Formula	18
Table 2. Conditional Density Recursion formulae for	
Bayesian Estimation	19
 Chapter III	
Table 1. Outline of a Gaussian-Sum Recursion Procedure	70
 Chapter IV	
Table 1. The Normalized Standard Deviation $2\left(\frac{P-1}{N}\right)^{1/2}$ as a	
function of P and N	82
Table 2. Monte Carlo Moments of Gaussian Generator	88
Table 3. Testing the Hypothesis of Gaussian Moments	88
Table 4. $\Pr(K_N < \lambda)$ from Massey	90
Table 5. Results of Kolmogorov Test	92
Table 6. Sampled Correlation Function	93
Table 7. Uncorrelatedness Test	94
Table A-1. Partition Requirements for m=36 bit random	
numbers	102
Table A-2. Partition Examples	103
Table A-3. Initial Sample Path - Sequence One	105
Table A-4. Initial Sample Path - Sequence Two	106

	<u>Page</u>
Table A-5. Repeat Characteristics of the Generator for each Sequence	107
Table A-6. FORTRAN-II Coding Examples	
Two-Piece Generator	108
Three-Piece Generator	109
Four-Piece Generator	110
Six-Piece Generator	111
Table A-7. FORTRAN-II Coding of Gaussian Generator	113
Chapter VI	
Table A-1. Monte-Carlo Performance of the Optimal and Linearized Predictors	142
Table A-2. Monte-Carlo Performance of Optimal and Linearized Filters	143
Table A-3. Monte-Carlo Confidence Intervals for Predictors	144
Table B-1. Monte-Carlo Average Sum Squared Error Performance for Predictors - Old Problem	149
Table B-2. Monte-Carlo Averaged Sum-Squared Error Performance for Predictors - New Problem	152
Chapter VII	
Table 1. Summary of Continuous Linearized Kalman-Bucy Filter	185
Table 2. Summary of Discrete Linearized Kalman-Bucy Filter	191
Table A-1. Confidence Intervals for the Linearized Filter	212
Table B-1. Numerical Values for the Computer Simulation	216

	<u>Page</u>
Table C-1. Monte Carlo Mod $2\pi$ Error Performance Data for the Cyclic Point-Mass Estimates	233
Table C-2. Monte Carlo Improvements - Cyclic Point-Mass over Phase-Locked Loop	234
Table C-3. Monte Carlo Difference Between Cyclic Point- Mass and Ideal Linear	234
Table C-4. Timing Estimates	238
Table C-5. $n/m$ Constant	239
Table C-6. $n$ Constant	240
Table C-7. $m$ Constant	241

## I. Introduction

This report is intended to serve as a record of specific experiments with feasible realizations of optimal nonlinear estimators. In addition, it is expected that future research along the lines described herein will continue to result in increased understanding of the behavior of optimal estimates and, possibly, in some guidelines for actual realizations in particular applications.

The underlying thread of continuity connecting all segments of this research is Bayes-Law (See Chapter II), the general solution to the discrete-time nonlinear estimation problem. Bayes-Law is in effect the discrete "representation theorem" (see Bucy and Joseph [7]). Some of the earliest attempts to realize Bayes Law on the digital computer involved orthogonal series representations of the conditional densities employing Gram-Charlier series [23], or Edgeworth expansion [22]. An early problem, however, concerned the tendency for truncated expansions to become predominantly negative resulting in unavoidable numerical instabilities.

In 1969 Bucy [4] proposed a point-mass approximation to density functions which involved a selection of important points on a "floating grid" and the centering of mass on the selected points. The point-mass approximation was of course always positive, easy to implement, and numerically stable. The computational burden was unfortunately prohibitive for high dimensional systems, however, and many short-cuts and simplifications were made by Bucy and Senne [10], [20] in order to make the computations tractable. A passive receiver problem was introduced by Bucy, Geesey, and Senne [6] to illustrate the concepts of point-mass approximation and the associated problems.

In an independent effort simultaneous to the above work, Alspach and Sorenson [2], [21] pioneered an approach based on a nonorthogonal series of Gaussians densities, originally chosen to be set down in such a way so as to minimize an  $L_p$  criterion, but subsequently to be determined via a simple approximation, again in order to provide a short-cut to the otherwise prohibitive computations.

In 1970 at the Nonlinear Estimation Symposium in San Diego Bucy and Senne [9] and Alspach and Sorenson [2] each described their respective approaches to the Bayes-Law computations, thereby providing the impetus for a sizable new interest in the computations associated with nonlinear estimation.

During the last two years, a multitude of topics related to Bayes-Law computation have emerged. Edison Tse [24] has noted a link between the previous two methods involving a Fourier transform translation theorem due to Wiener [26]. Julian Center [11] has observed the relationship between generalized least-squares projection and series expansions of density functions. Hecht [13] [14] has taken a much closer look at orthogonal polynomials - notably Gauss-Hermite expansions. Bucy, Merritt, and Miller [8], [18] have discussed hybrid solutions to reduce the serial computational burdens of Bayes Law by substituting the natural parallelisms of hybrid computers. Another promising approach to the approximation problem involving generalized splines has been studied by deFigueiredo and Jan [12], [15], while Weinert and Kailath [25] have been relating splines to least-squares approximation projection. Thus the subjects of numerical methods and optimal nonlinear estimation are now firmly entrenched.

Meanwhile, still another practical application of Bayesian estimation has recently been studied by Mallinckrodt, Bucy, and Cheng [17], by Hecht [14], by Bucy [5], and by Senne [19], and is reviewed in the current report. The new application involves demodulation of phase-modulated signals observed in additive white noise. Since the nominal engineering solution to such problems involves the well-known and reliable phase-locked loop, it appears that the demodulation problem will continue to provide an important comparison between moment series approximations and numerical density approximations.

Moment series approximations are, of course, the most commonly encountered nonlinear estimates in engineering practice today. The simplest moment approximation has been referred to by the names "extended" or "linearized" Kalman-Bucy filter [7]. The appeal of such methods is highly warranted in many problems, since the nonlinearities are not severe (i.e. they don't have arbitrarily large derivatives), and frequently the assumption of Gaussian noises is adequate. Whenever either or both of the "well-behaved" assumptions is false, however, considerable controversy has resulted. Some have advocated higher-order moment expansions, [3], others have proposed adaptive noise tracking techniques [16], or finite-memory filters [16], but generally nobody seems to ask the most fundamental question: What characteristics of the filtering problem have led to the demise of the simple first-order method? Or, equivalently, how would an optimal estimate behave in such a situation? The answers to these and other questions are directly addressed in the present report.

The existing organization of the report was necessitated due to time constraints, and although many different topics are addressed, there is occasionally some duplication. The attempt was to assemble a chronolog of the more significant results of the authors during the past three years into one source, thereby providing a focal point for subsequent research in the field. We apologize beforehand for any unavoidable difficulties for the reader caused by the presentation of the material. The global organization of the chapters is as follows:

Chapter II contains a simplified summary of a derivation of the principal Bayes-Law formulas used throughout the report. The presentation is taken principally from the dissertation of Hecht [14]. Chapter III provides a background for the various proposed numerical representations of the conditional densities. Covered in greatest detail are the orthogonal series, exemplified by Gauss-Hermite and Fourier, and the point-mass representation of Bucy and Senne [10]. Discussed in lesser detail are the nonorthogonal series (such as Gaussian sums) and generalized spline functions.

In Chapter IV the important topic of Monte Carlo simulation is treated in considerable depth. In particular, the subject of experimental confidence is treated in great detail and an example of the analysis is given involving the testing of the gaussian random number generator (which is realizable on any binary computer) to determine its statistical properties. Thus the reader is left with a complete understanding of the experimental methods used for this report.

Chapter V describes another important side-light of the current investigation - computer architecture. The concepts of parallelism and asynchronous computation are introduced and the Bayes estimation

problem is interpreted in light of parallel digital architecture. At first the "ideal" machine is postulated for computation of Bayes law. Then, as allowance is made for technical feasibility and current computer architecture, observations are made concerning efficient use of such structures as array processors (like Illiac IV), pipe-line machines (like CDC Star), look-ahead machines (like CDC 6600 or 7600), associative processors (like Goodyear's), and multi-processors (like the Burroughs D-Machines). Finally, some consideration is given to the currently available hybrid computer systems - examining their intrinsic parallelism.

Chapters VI and VII provide the details concerning the two examples studied in this work. The first example deals with a receive-only tracking system or a passive tracking receiver, which attempts to locate a target on the basis of bearing information imbedded in additive noise. The problem description is very similar to the Airborne Warning and Command System (AWACS), which is currently under contract development for the Air Force. The other example deals with phase demodulation, whereby a phase-modulated signal is observed in additive noise and it is desired to retrieve the original message process - at least modulo- $2\pi$ .

Chapter VIII contains a brief conclusion concerning the Bayes estimation research and indicates some paths for future developments.

The appendices provide documentation on some of the computer programs used and some unpublished technical papers relevant to the current research.

## References

- [ 1] D. L. Alspach, "A Bayesian Approximation Technique for Estimation and Control of Time Discrete Stochastic Systems," Ph.D. Dissertation, University of California, San Diego, 1970.
- [ 2] D. L. Alspach and H. W. Sorenson, "Approximation of Density Functions by a Sum of Gaussians for Nonlinear Bayesian Estimation," Proc. Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1970, 19-31.
- [ 3] R. W. Bass, V. D. Norum, and L. Schwartz, "Optimal Multichannel Non-Linear Filtering," J. Math. Anal. Appl. 16 (1966), 152-164.
- [ 4] R. S. Bucy, "Bayes Theorem and Digital Realizations for Non-Linear Filters," J. Astro. Sci. 17 (1969), 80-94.
- [ 5] R. S. Bucy, "Realization of Non-Linear Filters," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1971, 51-58.
- [ 6] R. S. Bucy, R. A. Geesey, and K. D. Senne, "Passive Receiver Design via Nonlinear Filtering Theory," Proc. Third Hawaii International Conf. on System Sciences, Vol I, 1970, 477-480.
- [ 7] R. S. Bucy and P. D. Joseph, Filtering for Stochastic Processes with Applications to Guidance, Wiley Interscience, New York, 1968.
- [ 8] R. S. Bucy, M. J. Merritt, and D. S. Miller, "Hybrid Computer Synthesis of Optimal Discrete Nonlinear Filters," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1971, 59-87.
- [ 9] R. S. Bucy and K. D. Senne, "Realization of Optimum Discrete-Time Nonlinear Estimators," Proc. Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1970, 6-17.
- [10] R. S. Bucy and K. D. Senne, "Digital Synthesis of Nonlinear Filters," Automatica 7 (1971), 287-298.
- [11] J. L. Center, "Practical Nonlinear Filtering of Discrete Observations by Generalized Least Squares Approximation of the Conditional probability Distribution," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1971, 88-99.
- [12] R. J. P. deFigueiredo and Y. G. Jan, "Spline Filters," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1971, 127-138.

## References (Cont)

- [13] C. Hecht, "Digital Realization of Non-Linear Filters," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1971, 152-158.
- [14] C. Hecht, "Synthesis and Realization of Nonlinear Filters," Ph.D. Dissertation, University of Southern California, 1972.
- [15] Y. G. Jan, Ph.D. Dissertation, Rice University, 1971.
- [16] A. H. Jazwinski, Stochastic Processes and Filtering Theory, Academic Press, New York, 1970.
- [17] A. J. Mallinckrodt, R. J. Bucy, and S. Y. Cheng, "Final Project Report for a Design Study for an Optimal Non-Linear Receiver/Demodulator," NASA Contract NAS5-10789, Goddard Space Flight Center, Maryland, 1970.
- [18] D. S. Miller, "Hybrid Synthesis of Optimal Discrete Nonlinear Filters," Ph.D. Dissertation, University of Southern California, 1971.
- [19] K. D. Senne, "Bayes Law Implementation: Optimal Discrete-Time Phase Estimation," Proc. SWIEEEO Conf., Dallas, April 1972.
- [20] K. D. Senne and R. S. Bucy, "Digital Realization of Optimal Discrete-Time Nonlinear Estimators", Proc. Fourth Annual Princeton Conf. on System Sciences, Princeton, March 1970, 280-284.
- [21] H. W. Sorenson and D. L. Alspach, "Recursive Bayesian Estimation using Gaussian Sums," Automatica, 7 (1971), 465-479.
- [22] H. W. Sorenson and A. R. Stubberud, "Non-Linear Filtering by Approximation of the A Posteriori Density," International J. Control, 8 (1968), 33-51.
- [23] K. Srinivasan, "State Estimation by Orthogonal Expansion of Probability Distributions," IEEE Trans. Auto. Control, AC-15 (1970), 3-10.
- [24] E. Tse, "Parallel Computation of the Conditional Mean State Estimate for Nonlinear Systems," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1971, 385-394.
- [25] H. L. Weinert and T. Kailath, "Recursive Spline Interpolation and Least-Squares Estimation," submitted to Amer. Math. Soc., 1971.
- [26] N. Wiener, The Fourier Integral and Certain of Its Applications, Cambridge, Cambridge University Press, 1933 (Also: New York, Dover, 1958).

## II. Bayesian Estimation: The Problem

Although the equations for Bayesian estimation are relatively well known, having been derived for example by Bucy [1], [2], a modified derivation is included in this chapter for the sake of introducing relevant notation and to make the present exposition as self-contained as possible. This presentation is taken from Hecht [3].

The discrete-time process and measurement equations may be written as

$$\underline{x}_n = \phi(\underline{x}_{n-1}) + \sigma_{n-1}(\underline{x}_{n-1})\underline{u}_{n-1} \quad (1)$$

$$\underline{x}_0 = \underline{c}$$

$$\underline{z}_n = \underline{h}(\underline{x}_n) + \underline{v}_n \quad (2)$$

Equation (1) represents a discrete time signal process with  $\underline{x}_n$  a sequence of  $d$ -dimensional random vectors; the subscript  $n$  refers to time.  $\phi(\underline{x}_{n-1})$  is a function from  $R^d$  to  $R^d$ ,  $\sigma(\underline{x}_{n-1})$  a function from  $R^d$  to  $d \times r$  matrices. The process  $\{\underline{u}_n\}_{n=1, \dots}$  is a set of independent  $r$ -dimensional random vectors with density  $p_{\underline{u}_n}(\underline{u})$ . The random vector  $\underline{c}$  is  $d$ -dimensional, independent of the  $\underline{u}_n$  process, and has density  $p_c(\underline{x})$ .

Equation (2) represents the observation process, with  $\underline{z}$  a sequence of  $s$ -dimensional random vectors,  $\underline{h}(\underline{x}_n)$  a function from  $R^d$  to  $R^s$  and  $\{\underline{v}_n\}_{n=1, \dots}$  a set of independent  $s$ -dimensional random vectors with density  $p_{\underline{v}_n}(\underline{v})$ , independent of  $\underline{c}$  and the  $\underline{u}_n$  process.

The filtering problem consists of determining the conditional density, given as

**Preceding page blank**

$$J_{n|t}(\underline{y}) d\underline{y} = P(\underline{x}_n \in d\underline{y} | \underline{z}_t = \underline{z}_t, \dots, \underline{z}_0 = \underline{z}_0) . \quad (3)$$

Results are stated in the same notation as Bucy [2]. The following notation will be used in discussing the derivation of the conditional densities.

Underlined lower-case Latin letters denote the name of random variables or random vectors, and related Greek letters represent the dummy argument associated with the density or distribution functions.

$p(\cdot)$  = probability density function

$P(\cdot)$  = probability distribution function

(The above functions are referred to briefly as "density" and "distribution" respectively.)

Thus, for example,

$p_{\underline{x}_n}(\underline{\xi}_n)$  = the density function of the random vector  $\underline{x}_n$ , with  $\underline{\xi}_n$  as dummy argument.

$p_{\underline{x}_n}(\underline{\xi}_n | \underline{z}_n = \underline{\zeta}_n)$  = the conditional density of the random vector  $\underline{x}_n$ , given the random vector  $\underline{z}_n$  has taken on the value  $\underline{\zeta}_n$ .

$p_{\underline{x}_n}(\underline{\xi}_n | \underline{z}_n = \underline{\zeta}_n)$  is a function of  $\underline{\xi}_n$  and  $\underline{\zeta}_n$ .

The above is frequently abbreviated to  $p_{\underline{x}_n}(\underline{\xi}_n | \underline{z}_n = \underline{\zeta}_n) = p_{\underline{x}_n}(\underline{\xi}_n | \underline{z}_n)$ , with the argument  $\underline{\zeta}_n$  implied.  $p_{\underline{x}_n, \underline{x}_{n-1}}(\underline{\xi}_n, \underline{\xi}_{n-1})$  = the joint density of the random vectors  $\underline{x}_n$  and  $\underline{x}_{n-1}$ .

Using the above notation (3) is

$$J_{n|t}(\underline{\xi}) d\underline{\xi} = p_{\underline{x}_n}(\underline{\xi}_n | \underline{z}_t = \underline{z}_t, \dots, \underline{z}_0 = \underline{z}_0) . \quad (4)$$

It is noted in Bucy's original work that the signal process is a Markov process with transition density

$$p_{x_{r+j}}(\xi_{r+j} | x_j = \xi_j) . \quad (5)$$

The required recursion relations for the conditional density  $J_{n|n}$  are given by the following equations.

$$J_{n|n}(\xi_n) = \frac{1}{K(n)} p_{v_n}[\xi_n - h(\xi_n)] \int d\xi p_{\sigma u_n}[\xi_n - \phi(\xi_{n-1})] J_{n-1|n-1}(\xi_{n-1}) d\xi_{n-1}, \quad (6)$$

with

$$J_{0|0}(\xi_0) = \frac{1}{K(0)} p_{v_0}[\xi_0 - h(\xi_0)] p_c(x), \quad (7)$$

$$K(0) = p_{z_0}(\xi_0), \quad (8)$$

$$K(n) = p_{z_n}(\xi_n | z_{n-1}, \dots, z_0), \quad (9)$$

$$J_{n+r|n}(\xi_{n+r}) = \int d\xi p_{x_{n+r}}(\xi_{n+r} | x_n = \xi_n) J_{n|n}(\xi_n) d\xi_n, \quad (10)$$

$$J_{n+1|n}(\xi_{n+1}) = \frac{1}{K(n)} \int d\xi p_{\sigma u_{n-1}}[\xi_{n+1} - \phi(\xi_n)] p_{v_n}[\xi_n - h(\xi_n)] J_{n|n-1}(\xi_n) d\xi_n \quad (11)$$

where previously undefined symbols have the following meaning:

$$\begin{aligned} \int d\xi (\cdot) d\xi &= \int d(\cdot) dx_1, \dots, dx_d, \\ &= d \text{ integrations} \end{aligned}$$

and  $p_{\sigma u_{n-1}}(\cdot)$  = the density of the  $d$ -dimensional random vector

$$\sigma(x_{n-1}) u_{n-1}.$$

The derivation of (6) through (11) follows:

$$J_{0|0}(\xi_0) d\xi_0 = p_{x_0}(\xi_0 | z_0 = \xi_0) = \frac{p_{z_0}(\xi_0 | x_0 = \xi_0) p_{x_0}(\xi_0)}{p_{z_0}(\xi_0)} \quad (12)$$

by Bayes rule.

$$\begin{aligned}
p_{z_0}(\xi_0 | x_0 = \xi_0) & \text{ corresponds to the distribution } P(z_0 \leq \xi_0 | x_0 = \xi_0) \\
& = P[h(x_0) + v_0 \leq \xi_0 | x_0 = \xi_0] \\
& = P[v_0 \leq \xi_0 - h(\xi_0)]
\end{aligned}$$

which corresponds to

$$p_{v_0}[\xi_0 - h(\xi_0)], \text{ or } p_{z_0}(\xi_0 | x_0 = \xi_0) = p_{v_0}[\xi_0 - h(\xi_0)]. \quad (13)$$

Substituting (13) into (12) gives

$$\begin{aligned}
p_{x_0}(\xi_0 | z_0 = \xi_0) & = \frac{1}{p_{z_0}(\xi_0)} p_{v_0}[\xi_0 - h(\xi_0)] p_{x_0}(\xi_0) \\
& = \frac{1}{K(0)} p_{v_0}[\xi_0 - h(\xi_0)] p_c(x)
\end{aligned}$$

which is (7). Next, using relations of conditional and joint densities,

$$\begin{aligned}
J_n | n (\xi_n) d\xi_n & = p_{x_n}(\xi_n | z_n, \dots, z_0) \\
& = \frac{p_{x_n, z_n, \dots, z_0}(\xi_n, \xi_n, \dots, \xi_0)}{p_{z_n, \dots, z_0}(\xi_n, \dots, \xi_0)} \\
& = \frac{p_{x_n, z_n}(\xi_n, \xi_n | z_{n-1}, \dots, z_0) p_{z_{n-1}, \dots, z_0}(\xi_{n-1}, \dots, \xi_0)}{p_{z_n}(\xi_n | z_{n-1}, \dots, z_0) p_{z_{n-1}, \dots, z_0}(\xi_{n-1}, \dots, \xi_0)} \\
& = \frac{1}{K(n)} p_{x_n, z_n}(\xi_n, \xi_n | z_{n-1}, \dots, z_0) \\
& = \frac{1}{K(n)} \int \int p_{x_n, z_n, x_{n-1}}(\xi_n, \xi_n, \xi_{n-1} | z_{n-1}, \dots, z_0) d\xi_{n-1}
\end{aligned}$$

Now operating on the integrand,

$$\begin{aligned}
& p_{x_n, z_n, x_{n-1}}(\xi_n, \zeta_n, \xi_{n-1}) \\
& = p_{z_n}(\zeta_n | x_n, x_{n-1}, z_{n-1}, \dots, z_0) p_{x_n, x_{n-1}}(\xi_n, \xi_{n-1} | z_{n-1}, \dots, z_0) \quad (15)
\end{aligned}$$

Next, by independence of the  $u_n$  and  $v_n$  processes and the Markov property,

$$p_{z_n}(\zeta_n | x_n, x_{n-1}, z_{n-1}, \dots, z_0) = p_{z_n}(\zeta_n | x_n) \quad (16)$$

which may be manipulated using the corresponding distribution function.

$$\begin{aligned}
P(z_n \leq \zeta_n | x_n = \xi_n) &= P[h(x_n) + v_n \leq \zeta_n | x_n] \\
&= P[h(\xi_n) + v_n \leq \zeta_n] = P[v_n \leq \zeta_n - h(\xi_n)] \quad (17)
\end{aligned}$$

with corresponding density  $p_{v_n}[\zeta_n - h(\xi_n)]$ , or

$$p_{z_n}(\zeta_n | x_n, x_{n-1}, z_{n-1}, \dots, z_0) = p_{v_n}[\zeta_n - h(\xi_n)] \quad (18)$$

Also,

$$\begin{aligned}
& p_{x_n, x_{n-1}}(\xi_n, \xi_{n-1} | z_{n-1}, \dots, z_0) \\
& = p_{x_n}(\xi_n | x_{n-1}, z_{n-1}, \dots, z_0) p_{x_{n-1}}(\xi_{n-1} | z_{n-1}, \dots, z_0) \\
& = p_{x_n}(\xi_n | x_{n-1}) J_{n-1|n-1}(\xi_{n-1}) \quad (19)
\end{aligned}$$

by independence of the  $u_n$  and  $v_n$  processes, the Markov property, and the definition of  $J_{n-1|n-1}(\xi_{n-1})$ . Again by manipulating the distribution function corresponding to  $p_{x_n}(\xi_n | x_{n-1})$ ;

$$\begin{aligned}
P(\underline{x}_n \leq \underline{\xi}_n | \underline{x}_{n-1}) &= P[\phi(\underline{x}_{n-1}) + \sigma(\underline{x}_{n-1}) \underline{u}_{n-1} \leq \underline{\xi}_n | \underline{x}_{n-1}] \\
&= P[\sigma(\underline{x}_{n-1}) \underline{u}_{n-1} \leq \underline{\xi}_n - \phi(\underline{x}_{n-1}) | \underline{x}_{n-1}] \\
&= P[\sigma(\underline{\xi}_{n-1}) \underline{u}_{n-1} \leq \underline{\xi}_n - \phi(\underline{\xi}_{n-1})] \quad (20)
\end{aligned}$$

the corresponding density is obtained,

$$p_{\underline{x}_n}(\underline{\xi}_n | \underline{x}_{n-1}) = p_{\sigma(\underline{x}_{n-1}) \underline{u}_{n-1}}[\underline{\xi}_n - \phi(\underline{\xi}_{n-1})] \quad (21)$$

Putting (18), (19), and (21) into (14),

$$\begin{aligned}
J_n | n(\underline{\xi}_n) d\underline{\xi}_n \\
= \frac{1}{K(n)} p_{\underline{v}_n}[\underline{\xi}_n - \underline{h}(\underline{\xi}_n)] \int d\underline{\sigma}(\underline{x}_{n-1}) \underline{u}_{n-1} [\underline{\xi}_n - \phi(\underline{\xi}_{n-1})] J_{n-1} | n-1(\underline{\xi}_{n-1}) d\underline{\xi}_{n-1} \quad (22)
\end{aligned}$$

which is (6).

Making the substitutions  $\underline{\xi}_n = \underline{y}$ ,  $\underline{\xi}_{n-1} = \underline{x}$ , assuming the  $\underline{u}_n$  and  $\underline{v}_n$  processes are zero mean gaussian with variances  $Q$  and  $R$ , and using the notation  $N(\underline{\xi}, \Lambda)$  for gaussian density with covariance  $\Lambda$  and argument  $\underline{\xi}$ , (22) becomes

$$J_n | n(\underline{y}) = \frac{1}{K(n)} N[\underline{\xi}_n - \underline{h}(\underline{y}), R] \int d\underline{\sigma}(\underline{x}) N(\underline{y} - \phi(\underline{x}), \sigma Q \sigma') J_{n-1} | n-1(\underline{x}) d\underline{x} \quad (23)$$

Equations (6), (22) and (23) are referred to as the conditional filter density update equations (for the gaussian special case).

Continuing with the development,

$$\begin{aligned}
J_{n+r} | n(\underline{\xi}_{n+r}) &= p_{\underline{x}_{n+r}}(\underline{\xi}_{n+r} | \underline{z}_n, \dots, \underline{z}_0) \\
&= \int d\underline{\sigma} p_{\underline{x}_{n+r}, \underline{x}_n}(\underline{\xi}_{n+r}, \underline{\xi}_n | \underline{z}_n, \dots, \underline{z}_0) d\underline{\xi}_n \\
&= \int d\underline{\sigma} p_{\underline{x}_{n+r}}(\underline{\xi}_{n+r} | \underline{x}_n, \underline{z}_n, \dots, \underline{z}_0) p_{\underline{x}_n}(\underline{\xi}_n | \underline{z}_n, \dots, \underline{z}_0) d\underline{\xi}_n \\
&= \int d\underline{\sigma} p_{\underline{x}_{n+r}}(\underline{\xi}_{n+r} | \underline{x}_n) p_{\underline{x}_n}(\underline{\xi}_n | \underline{z}_n, \dots, \underline{z}_0) d\underline{\xi}_n \quad (24)
\end{aligned}$$

by the Markov property.

$$J_{n+r|n}(\xi_{n+r}) = \int d\xi p_{x_{n+r}}(\xi_{n+r} | x_n) J_{n|n}(\xi_n) d\xi_n \quad (25)$$

which is (10).

Noting  $p_{x_{n+1}}(\xi_{n+1} | x_n) = p_{\sigma(x_n)u_n}[\xi_{n+1} - \phi(\xi_n)]$  from (21), and

substituting into (25),

$$J_{n+1|n}(\xi_{n+1}) = \int d\xi p_{\sigma(x_n)u_n}[\xi_{n+1} - \phi(\xi_n)] J_{n|n}(\xi_n) d\xi_n \quad (26)$$

then substituting (22) for  $J_{n|n}(\xi_n) d\xi_n$

$$\begin{aligned} J_{n+1|n}(\xi_{n+1}) &= \int d\xi p_{\sigma(x_n)u_n}[\xi_{n+1} - \phi(\xi_n)] \\ &\cdot \left\{ \frac{1}{K(n)} p_{v_n}[\xi_n - h(\xi_n)] \int d\xi p_{\sigma(x_{n-1})u_{n-1}}[\xi_n - \phi(\xi_{n-1})] J_{n-1|n-1}(\xi_{n-1}) d\xi_{n-1} \right\} d\xi_n \end{aligned}$$

using (21),

$$\begin{aligned} &= \int d\xi p_{\sigma(x_n)u_n}[\xi_{n+1} - \phi(\xi_n)] \frac{1}{K(n)} p_{v_n}[\xi_n - h(\xi_n)] \\ &\cdot \left\{ \int d\xi p_{x_n}(\xi_n | x_{n-1}) J_{n-1|n-1}(\xi_{n-1}) d\xi_{n-1} \right\} d\xi_n \\ &= \frac{1}{K(n)} \int d\xi p_{\sigma(x_n)u_n}[\xi_{n+1} - \phi(\xi_n)] p_{v_n}[\xi_n - h(\xi_n)] J_{n|n-1}(\xi_n) d\xi_n \quad (27) \end{aligned}$$

by again using (25). Equations (27) and (11) are identical.

By substituting  $y = \xi_{n+1}$ ,  $x = \xi_n$  and assuming the same gaussian conditions following (22), (27) becomes

$$J_{n+1|n}(y) = \frac{1}{K(n)} \int dN[y, \sigma Q \sigma'] N[z - h(x), R] J_{n|n-1}(x) dx \quad (28)$$

Equations (11), (27), and (28) are referred to as the conditional one-step predictor density update equations.

Relations between the one-step predictor and the filter density update equations may also be written as follows.

Use (26) for the one-step predictor in the filter equation, (22), to obtain

$$J_{n|n}(\xi_n) = \frac{1}{K(n)} p_{v_n}[\xi_n - h(\xi_n)] J_{n|n-1}(\xi_n), \quad (29)$$

and use (29) in (27) to obtain

$$J_{n+1|n}(\xi_{n+1}) = \frac{1}{K(n)} \int d\phi_{ou_n} [\xi_{n+1} - \phi(\xi_n)] J_{n|n}(\xi_n) d\xi_n \quad (30)$$

Tables 1 and 2 summarize the above results.

For completeness the smoothing filter is also included, although these results were not used in this research.

$$\begin{aligned} p_{x_{n-r}}(\xi_{n-r} | z_n, \dots, z_0) &= J_{n-r|n}(\xi_{n-r}) \\ &= \frac{p_{x_{n-r}, z_n, \dots, z_0}(\xi_{n-r}, \xi_n, \dots, \xi_0)}{p_{z_n, \dots, z_{n-r+1}}(\xi_n, \dots, \xi_{n-r+1} | z_{n-r}, \dots, z_0) p_{z_{n-r}, \dots, z_0}(\xi_{n-r}, \dots, \xi_0)} \end{aligned} \quad (31)$$

The first term in the denominator is abbreviated  $c(n, r)$ .

$$\begin{aligned} p_{x_{n-r}}(\xi_{n-r} | z_n, \dots, z_0) &= \frac{1}{c(n, r)} \int \int \\ &\cdot \frac{p_{x_{n-r}, z_n, \dots, z_0, x_n, \dots, x_{n-r+1}}(\xi_{n-r}, \xi_n, \dots, \xi_0, \xi_n, \dots, \xi_{n-r+1})}{p_{z_{n-r}, \dots, z_0}(\xi_{n-r}, \dots, \xi_0)} d\xi_n \dots d\xi_{n-r} \\ &= \frac{1}{c(n, r)} \int \int \frac{p_{x_{n-r}, z_{n-r}, \dots, z_0}(\xi_{n-r}, \xi_{n-r}, \dots, \xi_0)}{p_{z_{n-r}, \dots, z_0}(\xi_{n-r}, \dots, \xi_0)} \\ &\cdot p_{x_n, \dots, x_{n-r+1}, z_n, \dots, z_{n-r+1}}(\xi_n, \dots, \xi_{n-r+1}, \xi_n, \dots, \xi_{n-r+1} | x_{n-r} = \xi_{n-r}) \\ &\cdot d\xi_n \dots d\xi_{n-r+1} \end{aligned}$$

Using

$$\begin{aligned}
 & p_{x_{n-r}, z_{n-r}, \dots, z_0}(\xi_{n-r}, \xi_{n-r}, \dots, \xi_0) \\
 &= p_{x_{n-r}}(\xi_{n-r} | z_{n-r}, \dots, z_0) p_{z_{n-r}, \dots, z_0}(\xi_{n-r}, \dots, \xi_0) \\
 &= J_{n-r|n-r}(\xi_{n-r}) p_{z_{n-r}, \dots, z_0}(\xi_{n-r}, \dots, \xi_0) \\
 & p_{x_{n-r}}(\xi_{n-r} | z_n, \dots, z_0) = J_{n-r|n}(\xi_{n-r}) \\
 &= \frac{1}{c(n, r)} \int \int J_{n-r|n-r}(\xi_{n-r}) \\
 & \quad \cdot p_{x_n, \dots, x_{n-r+1}, z_n, \dots, z_{n-r+1}}(\xi_n, \dots, \xi_{n-r+1}, \xi_n, \dots, \xi_{n-r+1} | x_{n-r}) \\
 & \quad \cdot d\xi_n \dots d\xi_{n-r+1} \\
 &= \frac{1}{c(n, r)} \int \int J_{n-r|n-r}(\xi_{n-r}) \prod_{j=n-r+1}^n p_{v_j}[\xi_j - h(\xi_j)] \\
 & \quad \cdot p_{ou}[\xi_j - \phi(\xi_{j-1})] d\xi_n \dots d\xi_{n-r+1}
 \end{aligned}$$

Table 1.

## Model for Bayes Rule Conditional Density Recursion Formula

$$\underline{x}_n = \phi_n(\underline{x}_{n-1}) + \sigma_{n-1}(\underline{x}_{n-1})\underline{u}_{n-1} \quad (1)$$

$$\begin{aligned} \underline{x}_0 &= \underline{c} \\ \underline{z}_n &= \underline{h}_n(\underline{x}_n) + \underline{v}_n \end{aligned} \quad (2)$$

Notation

$\underline{x}_n$	= a sequence of d-dimensional random vectors
$n$	= time index
$\phi_n(\underline{x}_{n-1})$	= a function from $R^d$ to $R^d$
$\sigma_{n-1}(\underline{x}_{n-1})$	= a function from $R^d$ to dxr matrices
$\{\underline{u}_n\}$	= a process of independent r-dimensional random vectors with density $p_{u_n}(\omega)$
$\underline{c}$	= a d-dimensional random vector independent of the $\underline{u}_n$ process and having density $p_c(\underline{x})$
$\underline{z}_n$	= a sequence of s-dimensional random vectors
$\underline{h}_n(\underline{x}_n)$	= a function from $R^d$ to $R^s$
$\{\underline{v}_n\}$	= a process of independent s-dimensional random vectors with density $p_{v_n}(\phi)$ , independent of $\underline{c}$ and the $\underline{u}_n$ processes

Table 2.

## Conditional Density Recursion Formulae for Bayesian Estimation

Recurrence Relation

$$\begin{aligned}
J_{n-1|n-1}(y) \rightarrow J_{n|n}(y) &= \frac{1}{k(n)} p_v[z_n - \bar{h}(y)] \int d p_{\sigma_u} [\bar{y} - \bar{\phi}(\bar{x})] J_{n-1|n-1}(\bar{x}) d\bar{x} \\
&= \frac{1}{k(n)} N[\bar{z}_n - \bar{h}(y), R] \int d N[\bar{y} - \bar{\phi}(\bar{x}), \sigma Q \sigma'] J_{n-1|n-1}(\bar{x}) d\bar{x} \\
J_{n|n-1}(y) \rightarrow J_{n|n}(y) &= \frac{1}{k(n)} p_v[z_n - \bar{h}(y)] J_{n|n-1}(y) \\
&= \frac{1}{k(n)} \cdot N[\bar{z}_n - \bar{h}(y), R] J_{n|n-1}(y) \\
J_{n|n-1}(y) \rightarrow J_{n+1|n}(y) &= \frac{1}{k(n)} \int d p_{\sigma_u} [\bar{y} - \bar{\phi}(\bar{x})] p_v[z_{n+1} - \bar{h}(\bar{x})] J_{n|n-1}(\bar{x}) d\bar{x} \\
&= \frac{1}{k(n)} \int d N[\bar{y} - \bar{\phi}(\bar{x}), \sigma Q \sigma'] N[\bar{z}_{n+1} - \bar{h}(\bar{x}), R] J_{n|n-1}(\bar{x}) d\bar{x} \\
J_{n|n}(y) \rightarrow J_{n+1|n}(y) &= \frac{1}{k(n)} \int d p_{\sigma_u} [\bar{y} - \bar{\phi}(\bar{x})] J_{n|n}(\bar{x}) d\bar{x} \\
&= \frac{1}{k(n)} \int d N[\bar{y} - \bar{\phi}(\bar{x})] J_{n|n}(\bar{x}) d\bar{x}
\end{aligned}$$

Table 2. Continued

Note: The index  $n$  has been omitted in some places for notational simplicity.

Notation

$J_{a|b}(y)$  = conditional density of the random vector  $\underline{x}_a$  given observation vectors  $\underline{z}_0, \dots, \underline{z}_b$

$\kappa(n)$  = a normalizing constant to make  $J(\cdot)$  a density function

$\int d\int(\cdot) d\bar{x}$  =  $d$ -fold integration with respect to the variables  $x_i$ , components of the vector  $\underline{x}$ .

$P_{\sigma_u}(\cdot)$  = density of the  $r$ -dimensional random vector  $\sigma_{n-1}(\underline{x}_{n-1})_{u_{n-1}}$

$N(\underline{\xi}, \Lambda)$  = gaussian density with covariance  $\Lambda$  and argument  $\underline{\xi}$

## References

- [1] R.S. Bucy, "Bayes Theorem and Digital Realizations for Nonlinear Filters," J. Astro. Sci. 17 (1969), 80-94.
- [2] R.S. Bucy and P.D. Joseph, Filtering for Stochastic Processes with Applications to Guidance, Wiley Interscience, New York, 1968.
- [3] C. Hecht, "Synthesis and Realization of Nonlinear Filters," Ph.D. Dissertation, University of Southern California, 1972.

### III Finite Dimensional Approximations

#### A. Introduction

The formal Bayes-Law calculations, reviewed in detail in the previous chapter, are functionally simple but their implications to computation are formidable. To begin with, the representation of densities is in general an infinite-dimensional problem. Although there are many examples of density families characterized by a finite number of parameters, the Bayes-law computation rarely reproduces another member of the same family. The most well-known exception is the Gaussian family, which is reproduced under linear transformations, resulting in the widely used Kalman-Bucy filter, which optimally describes the Bayes-law computation in terms of linear differential (difference) equations for the mean and covariance of the Gaussian conditional densities, provided that the underlying physical system is described by linear differential (difference) equations with additive Gaussian inhomogeneities. If the physical plant is not linear or the disturbances are not Gaussian, however, the Bayes-law rarely leads to a reproducing family of densities. Thus an improvised or unnatural parameterization of the densities is required in order to implement Bayes-law and, in general, an infinite number of parameters is required for exact representation. The numerical approximation problem, then is simply stated: how do we choose an appropriate (finite) subset of the parameters to represent a given collection of conditional densities? Since the answer to this fundamental question depends heavily on the densities in question, it turns out that little can be said about the applicability of a given approximation without discussing specific example problems. Even for a given problem there

**Preceding page blank**

may be several different appropriate numerical approximation schemes, depending upon the problem parameters. Examples of these dependences will be discussed in Chapters VI and VII.

In the present paper we will discuss some representatives from several categories of density approximations as well as their Bayes-law implementations and associated difficulties. First, we discuss orthogonal series, covering a candidate with unbounded support (Gauss-Hermite polynomials - Section B) and also a candidate with compact support (Trigonometric series - Section E).

Next, we discuss an approach involving nonorthogonal polynomials which is intended to provide positive density approximations for any finite number of terms (Section D). Thirdly, we discuss an intuitively simple approach to density approximation involving point masses (Section C), suitably distributed so that most of the probability is adequately covered by a small number of discrete points. Finally, we describe a relatively recent additional approach to the problem using numerical spline functions (Section E). The presentation of this paper is intended only to be representative and not exhaustive, since there are many approaches to numerical approximation which have yet to be considered.

#### B. Orthogonal Series

The general theory of orthogonal series may be found in many places in the literature (see, for example, [10]). In this section we intend only to provide two contrasting examples, whereby we illustrate the significance of the type of state-space required for the application at hand. If the state vector can take on all values in  $\mathbb{R}^n$  with positive probability, then orthogonal polynomials must be used to provide the necessary approximation. On the other hand, if

the state-space actually is or can be approximated as compact then a periodic function (such as trigonometric) might profitably be employed as a basis for an orthogonal series. It may happen that a given problem may be interpreted in either way (see, for example, Chapter VII), so that more than one orthogonal series may be appropriate, depending on the performance desired.

### 1. Least Square Polynomial Approximation, Scalar Case

The theory given here follows Hildebrand [10], and is only an outline giving key results. Detailed proofs may be found in the reference.

We wish to approximate a function  $f(x)$  with a series of polynomials  $y(x)$ , as follows:

$$\hat{f}(x) \approx y(x) = \sum_{k=0}^n a_k \phi_k(x) \quad (1)$$

where  $\phi_0(x), \dots, \phi_n(x)$  are the required polynomial functions. The approximation is to be the best in the sense that

$$\begin{aligned} & \int_a^b w(x) [f(x) - y(x)]^2 dx \\ &= \int_a^b w(x) \left[ f(x) - \sum_{k=0}^n a_k \phi_k(x) \right]^2 dx = \text{Minimum} \end{aligned} \quad (2)$$

with  $w(x)$  a specified weighting function which is assumed non-negative on the interval  $(a, b)$ .

Equation (2) imposes the condition on the coefficients  $a_k$ ,

$$\frac{\partial}{\partial a_r} \int_a^b w(x) [f(x) - \sum_{k=0}^n a_k \phi_k(x)]^2 dx = 0 \quad (r=0,1,\dots,n) \quad (3)$$

from which

$$\sum_{k=0}^n a_k \int_a^b w(x) \phi_r(x) \phi_k(x) dx = \int_a^b w(x) \phi_r(x) f(x) dx \quad (r=0,1,\dots,n) \quad (4)$$

The coordinate functions are chosen to be orthogonal to each other over the interval  $(a,b)$  with respect to the weighting function  $w(x)$ .

$$\int_a^b w(x) \phi_r(x) \phi_k(x) dx = 0 \quad r \neq k \quad (5)$$

The "uncoupled" equations reduce to (omitting the argument  $x$ )

$$a_r \int_a^b w \phi_r^2 dx = \int_a^b w \phi_r f dx$$

or

$$a_r = \frac{\int_a^b w \phi_r f dx}{\int_a^b w \phi_r^2 dx} \quad (6)$$

To construct the polynomial functions  $\phi_0(x), \phi_1(x), \dots, \phi_r(x)$ , it is required that the polynomial  $\phi_r(x)$  be orthogonal to all polynomials of degree inferior to  $r$ , over the interval  $(a,b)$  with respect to the weighting function  $w(x)$ .

$$\int_a^b w(x) \phi_r(x) q_{r-1}(x) dx = 0 \quad (7)$$

where  $q_{r-1}$  is an arbitrary polynomial of degree  $r-1$  or less. The notation is introduced

$$w(x) \phi_r(x) = \frac{d^r U_r(x)}{dx^r} = l_r^{(r)}(x) \quad (8)$$

so that (7) becomes

$$\int_a^b U_r^{(r)}(x) q_{r-1}(x) dx = 0$$

After  $r$  integrations by parts

$$[U_r^{(r-1)} q_{r-1} - U_r^{(r-2)} q_{r-1}' + \dots + (-1)^{r-1} U_r q_{r-1}^{(r-1)}]_a^b = 0 \quad (9)$$

The requirement that  $\phi_r(x)$  be a polynomial of degree  $r$  implies that  $U_r(x)$ , from (8), satisfy the differential equation

$$\frac{d^{r+1}}{dx^{r+1}} \left[ \frac{1}{w(x)} \frac{d^r U_r(x)}{dx^r} \right] = 0 \quad (10)$$

in  $(a, b)$  whereas the requirement (9) be satisfied for any values of  $q_{r-1}(a)$ ,  $q_{r-1}(b)$ ,  $q'_{r-1}(a)$ ,  $q'_{r-1}(b)$ , etc. leads to the boundary conditions

$$\left. \begin{aligned} U_r(a) &= U'_r(a) = U''_r(a) = \dots = U^{(r-1)}_r(a) = 0 \\ U_r(b) &= U'_r(b) = \dots = U^{(r-1)}_r(b) = 0 \end{aligned} \right\} \quad (11)$$

For each integer  $r$ , a solution of (10) which satisfies the boundary conditions (11), is the  $r^{\text{th}}$  member of the set of polynomial functions,  $\phi_r(x)$ , given by

$$\phi_r(x) = \frac{1}{w(x)} \frac{d^r U_r(x)}{dx^r} \quad (12)$$

The numerator to compute the coefficient  $a_r$ , given by (6), is a function of  $f(x)$ . The denominator, designated  $\gamma_r$ , is independent of  $f$  and need be computed only once.

$$\gamma_r = \int_a^b w(x) \phi_r^2(x) dx = (-1)^r r! A_r \int_a^b U_r(x) dx \quad (13)$$

where  $A_r$  is the leading coefficient of  $\phi_r(x)$ .

$$\phi_r(x) = A_r x^r + A_{r-1} x^{r-1} + \dots + A_0$$

It is shown, for use in the integration formulas, that if  $w(x)$  does not change sign in  $(a,b)$ , the polynomial  $\phi_r(x)$  possesses  $r$  distinct real zeros, all of which lie in the interval  $(a,b)$ .

For application to the problem of Chapter VII of this paper we want to approximate

$$f(x) = y(x) \approx v(x) \sum_{r=0}^n b_r \phi_r(x) \quad (14)$$

such that

$$\int_a^b w(x) \left[ \frac{f(x)}{v(x)} - \sum_{r=0}^n b_r \phi_r(x) \right]^2 dx = \text{minimum} \quad (15)$$

which leads to the result

$$b_r = \frac{1}{\gamma_r} \int_a^b \frac{w}{v} f \phi_r dx \quad (16)$$

which is equivalent to minimizing the squared error  $(f-y)^2$  with respect to the weighting function  $\frac{w}{v^2}$ . In the application we let

$$w = v = e^{-\alpha^2 x^2} \quad (17)$$

and the interval  $(a,b)$  is  $(-\infty, \infty)$ , which leads to the Hermite formulas.

For the above choice of  $w(x)$

$$\phi_r(x) = e^{\alpha^2 x^2} \frac{d^r U_r}{dx^r}$$

where  $U_r$  satisfies (equation (10))

$$\frac{d^{r+1}}{dx} \left[ e^{\alpha^2 x^2} \frac{d^r U_r}{dx^r} \right] = 0 \quad (18)$$

and from the boundary conditions, (11) requires  $U_r$  and its first  $r-1$  derivatives to tend to zero as  $x \rightarrow \pm\infty$

The function

$$U_r(x) = C_r e^{-\alpha^2 x^2} \quad (19)$$

has the property that its  $r^{\text{th}}$  derivative is the product of itself and a polynomial of degree  $r$ . It therefore satisfies (18) and the boundary conditions.

$$\phi_r(x) = C_r e^{\alpha^2 x^2} \frac{d^r}{dx^r} (e^{-\alpha^2 x^2}) \quad (20)$$

The Hermite polynomial of degree  $r$  is defined by taking

$$C_r = (-1)^r \text{ and } \alpha^2 = 1.$$

$$H_r(x) = (-1)^r e^{x^2} \frac{d^r}{dx^r} (e^{-x^2}). \quad (21)$$

For

$$C_r = (-\alpha)^{-r}$$

$$\phi_r(x) = H_r(\alpha x) \quad (22)$$

the Hermite polynomials are determined from the recurrence formula

$$H_0(x) = 1 \quad ,$$

$$H_1(x) = 2x \quad ,$$

and

$$H_{r+1}(x) = 2xH_r(x) - 2rH_{r-1}(x) \quad . \quad (23)$$

Equation (14) takes the form

$$y(x) = e^{-\alpha^2 x^2} \sum_{r=0}^n b_r H_r(\alpha x) \quad , \quad (24)$$

with  $\alpha_r$  and  $b_r$  (from (13) and (16)) given by

$$\gamma_r = \frac{2^r r!}{\alpha} \sqrt{\pi} \quad (25)$$

$$b_r = \frac{\alpha}{2^r r! \sqrt{\pi}} \int_{-\infty}^{\infty} f(x) H_r(\alpha x) dx$$

by using (23) for  $A_r$  and (19) for  $U_r$ , with  $C_r$  as given above.

## 2. Least Square Polynomial Approximation, Multi-dimensional Case

We extend the above theory to multi-dimension functions. To be specific, and in accordance with the requirements of this paper, the theory is shown for a function of two variables. Generalization to higher dimension is straightforward (although messy, requiring double and triple indices to avoid awkward expressions).

The approximation is

$$f(x_1 x_2) \approx y(x_1 x_2) = \sum_{k_1=0}^{m_1} \sum_{k_2=0}^{m_2} a_{k_1 k_2} \phi_{k_1}(x_1) \phi_{k_2}(x_2) \quad (26)$$

with the requirement

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} w_1(x_1) w_2(x_2) [f(x_1 x_2) - y(x_1 x_2)]^2 dx_1 dx_2 = \text{minimum} \quad (27)$$

Differentiating with respect to  $a_{k_1 k_2}$  and setting the result to zero leads to

$$\begin{aligned} & \sum_{k_1=0}^{m_1} \sum_{k_2=0}^{m_2} a_{k_1 k_2} \int_{a_1}^{b_1} \int_{a_2}^{b_2} w_1(x_1) w_2(x_2) \phi_{r_1}(x_1) \phi_{r_2}(x_2) \phi_{k_1}(x_1) \phi_{k_2}(x_2) dx_1 dx_2 \\ &= \int_{a_1}^{b_1} \int_{a_2}^{b_2} w_1(x_1) w_2(x_2) \phi_{r_1}(x_1) \phi_{r_2}(x_2) f(x_1 x_2) dx_1 dx_2 \end{aligned} \quad (28)$$

Using orthogonality properties identical to those for the scalar case

$$\begin{aligned} \int_{a_1}^{b_1} w_1 \phi_{k_1} \phi_{r_1} dx_1 &= 0 & k_1 \neq r_1 \\ \int_{a_2}^{b_2} w_2 \phi_{k_2} \phi_{r_2} dx_2 &= 0 & k_2 \neq r_2 \end{aligned}$$

gives for the coefficients

$$a_{r_1 r_2} = \frac{\int_{a_1}^{b_1} \int_{a_2}^{b_2} w_1 w_2 \phi_{r_1} \phi_{r_2} f dx_1 dx_2}{\int_{a_1}^{b_1} w_1^2 \phi_{r_1}^2 dx_1 \int_{a_2}^{b_2} w_2^2 \phi_{r_2}^2 dx_2} \quad (29)$$

( $r_1 r_2 = 00, 01, 10, 11, 02, \dots, m_1 m_2$ )

The polynomial functions are the same ones used for the scalar case. The denominator of (29) is given by  $\gamma_{r_1} \gamma_{r_2}$  where  $\gamma_{r_i} = \gamma_{r_i}$  as given by (13).

The two-dimensional approximation that is needed is

$$f(x_1 x_2) \approx y(x_1 x_2) = v_1(x_1) v_2(x_2) \sum_{r_1=0}^{m_1} \sum_{r_2=0}^{m_2} b_{r_1 r_2} \phi_{r_1}(x_1) \phi_{r_2}(x_2)$$

where, analogous to the scalar case,

$$b_{r_1 r_2} = \frac{1}{\gamma_{r_1} \gamma_{r_2}} \int_{a_1}^{b_1} \int_{a_2}^{b_2} \frac{w_1 w_2}{v_1 v_2} f \phi_{r_1} \phi_{r_2} dx_1 dx_2 \quad (30)$$

Letting

$$w_1 = v_1 = e^{-\alpha_1^2 x_1^2}$$

$$w_2 = v_2 = e^{-\alpha_2^2 x_2^2}$$

$$a_1 = a_2 = -\infty$$

$$b_1 = b_2 = \infty$$

and using the results of the scalar case,

$$y(x_1, x_2) = e^{-\alpha_1^2 x_1^2} e^{-\alpha_2^2 x_2^2} \sum_{r_1=0}^{m_1} \sum_{r_2=0}^{m_2} b_{r_1 r_2} H_{r_1}(\alpha_1 x_1) H_{r_2}(\alpha_2 x_2) \quad (31)$$

$$(-\infty < x_1 < \infty, -\infty < x_2 < \infty)$$

$$b_{r_1 r_2} = \frac{\alpha_1 \alpha_2}{2^{r_1} r_1! 2^{r_2} r_2!} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) H_{r_1}(\alpha_1 x_1) H_{r_2}(\alpha_2 x_2) dx_1 dx_2 \quad (32)$$

### 3. Gauss-Hermite Integration

The results given here are developed in Hildebrand [9] for integration of a function of one variable. A simple extension is made here for functions of two variables. The parts of the theory presented are limited to those needed to explain the methods used.

The values of a function  $f(x)$  are assumed known at  $m$  points,  $x=x_1, x_2, \dots, x_m$ . If the value of the derivative,  $f'(x)$ , is known at the same points, a polynomial of degree  $2m-1$  can be constructed to agree with  $f(x)$  at the  $m$  points. An integration formula, using Lagrangian interpolation, would give a numerical integration equation with a degree of precision of  $(2m-1)^*$ . Use of Gaussian integration requires the  $m$  points to be selected in a certain way, but leads to formulas which have the same degree of precision without the requirement for knowledge of the derivatives at the  $m$  points.

---

\*An integration formula which yields exact results when  $f(x)$  is a polynomial of degree  $r$  or less, but fails to give exact results for at least one polynomial of degree  $r+1$ , possess a degree of precision of  $r$ . (Hildebrand [9]).

Use is made of the auxiliary functions

$$\pi(x) = (x-x_1)(x-x_2) \dots (x-x_m) \quad (33)$$

$$l_i(x) = \frac{\pi(x)}{(x-x_i)\pi'(x_i)} \quad (34)$$

with the properties

$$\pi(x_j) = 0$$

$$l_i(x_j) = \delta_{ij} \quad (\text{Kronecker delta})$$

Let  $y(x)$  be a polynomial of degree  $2m-1$ , which agrees with  $f$  and  $f'$  at the  $m$  points. It can be expressed in the form

$$y(x) = \sum_{k=1}^m h_k(x)f(x_k) + \sum_{k=1}^m \bar{h}_k f'(x_k) \quad (35)$$

where  $h_i$  and  $\bar{h}_i$  are polynomials of maximum degree  $2m-1$ . Using the properties of the auxiliary functions the polynomials  $h_i$  and  $\bar{h}_i$  are given by

$$\begin{aligned} h_i(x) &= [1-2l_i'(x)(x-x_i)][l_i(x)]^2 \\ \bar{h}_i(x) &= (x-x_i)[l_i'(x)]^2 \end{aligned} \quad (36)$$

Integration of  $f(x)$  times a weighting function  $w(x)$ , using the approximation  $y(x)$ , Equation (35) gives

$$\int_a^b w(x) f(x) dx = \sum_{k=1}^m H_k f(x_k) + \sum_{k=1}^m \bar{H}_k f'(x_k) + E \quad (37)$$

with the weighting coefficients defined by

$$H_i = \int_a^b w h_i dx \quad (38)$$

$$\bar{H}_i = \int_a^b w \bar{h}_i dx \quad (39)$$

and the error

$$E = \frac{f^{(2m)}(y)}{(2m)!} \int_a^b w(x) [\pi(x)]^2 dx \quad (40)$$

with  $y$  some point on the interval  $(a,b)$ .

Using (33), (34) and (36), Equation (39) can be expressed as

$$\bar{H}_i = \frac{1}{\pi'(x_i)} \int_a^b w(x) \pi(x) \ell_i(x) dx \quad (41)$$

so that  $\bar{H}_i$  will vanish if  $\pi(x)$  is orthogonal to  $\ell_i(x)$  over  $(a,b)$  relative to the weighting function  $w(x)$ . The points  $x_1, \dots, x_m$  are the  $m$  zeros of  $\pi(x)$ . Each  $\ell_i(x)$  is a polynomial of degree  $m-1$ , so the requirement that  $\pi(x)$  be orthogonal to all polynomials of degree inferior to  $m$  is a sufficient condition. (It is also shown to be a necessary condition). The orthogonality requirement is identical to the requirement given by (7), and therefore defines the same types of polynomials, normalized to make the leading coefficient equal to one (Equation (33)). The  $m$  zeros of  $\pi(x)$  are real, distinct, and are in the interval  $(a,b)$ .

A substantial amount of manipulation is required to determine, explicitly, the weights  $H_i$  defined by (38). With  $A_m$  the leading coefficient of  $\phi_m(x)$  (Equation (12) and  $\gamma_m$  from (13), the following two equivalent forms are determined:

$$H_i = \frac{-A_{m+1}\gamma_m}{A_m\phi'_m(x_i)\phi_{m+1}(x_i)} \quad (42)$$

$$H_i = \frac{A_m\gamma_{m-1}}{A_{m-1}\phi'_m(x_i)\phi_{m-1}(x_i)} \quad (43)$$

For Gauss-Hermite integration  $w(x)=e^{-x^2}$  over the interval  $(-\infty, \infty)$ , which results in

$$\pi(x) = \frac{1}{A_m} H_m(x) \quad (44)$$

with  $H_m$  the  $m^{\text{th}}$  Hermite polynomial (Equation (23)), and where

$$A_m = 2^m \quad (45)$$

and

$$\gamma_m = \sqrt{\pi} 2^m m! \quad (46)$$

from (25) with  $\alpha=1$ .

The coefficient  $H_i$  can be determined from either (42) or (43) with  $\phi_r$  replaced by the Hermite polynomial  $H_r$ , or equivalently using a relation of the Hermite polynomial and its derivative

$$H_i = \frac{2^{m+1} m! \sqrt{\pi}}{[H_{m+1}(x_i)]^2} \quad (47)$$

In summary the Gauss-Hermite formula is of the form

$$\int_{-\infty}^{\infty} e^{-x^2} f(x) dx = \sum_{k=1}^m H_k f(x_k) + E \quad (48)$$

where  $x_i$  is the  $i^{\text{th}}$  zero of  $H_m(x)$  and using (40) and  $\gamma_r$  as defined by (25)

$$E = \frac{m! \sqrt{\pi}}{2^m (2m)!} f^{(2m)}(\xi) \quad (49)$$

with  $\xi$  some point in  $(-\infty, \infty)$ .

An extension to functions of more than one variable is made following the scalar method and using the extension technique previously described for the polynomial approximation, to obtain

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-x_1^2} e^{-x_2^2} f(x_1, x_2) dx_1 dx_2 = \sum_{k_1=1}^{m_1} \sum_{k_2=1}^{m_2} H_{k_1} H_{k_2} f(x_{1k_1}, x_{2k_2}) \quad (50)$$

with the weights and zeros as defined for one dimensional integrations.

#### 4. Application of Polynomial Expansions to a Two-Dimensional Filtering Problem

In this section we will develop the equations of the previous section for a particular example, which we intend to discuss later in Chapter VII in great detail. The appropriate equations describing the Bayes-Law update for the phase demodulation problem (see Chapter VII) are repeated as

$$J_{n+1|n} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = c_1 \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2q\Delta} (y_2 - x_2)^2 \right\} J_{n|n} \begin{pmatrix} y_1 - x_1 \Delta \\ x_2 \end{pmatrix} dx_2, \quad (51)$$

and

$$J_{n|n} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = c_2 \exp \left\{ -\frac{\Delta}{2r} \left[ \left( z_1 - \cos y_1 \right)^2 + \left( z_2 - \sin y_1 \right)^2 \right] \right\} J_{n|n-1} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \quad (52)$$

Beginning from these equations we now will develop a Hermite recursion formula. We assume a set of quantities are available at a particular time which completely characterizes the conditional density function, and demonstrate how these same quantities are computed for the following time increment.

1. From the prior stage we have the means, covariance matrix, characteristic values and vectors of the covariance matrix, and a series expansion of the conditional density in characteristic vector coordinates in terms of Hermite polynomials.

2. The series form of the density is put into Equation (51) and then one takes the Fourier transform of both sides.
3. Rework the result of Step 2 to take the inverse Fourier transform to obtain  $J_{n+1|n}$ .
4. Multiply  $J_{n|n-1}$  by the observation process, Equation (52), to obtain  $J_{n|n}$ .
5. Compute moments for means, covariance matrix, and for the coefficients of the new expansion.

The above outline is illustrated in Figure 1. In the figure we see two types of operations, analytic and computational. The analytic operations, outlined in the above steps, enable one to compute  $J_{n|n}$  recursively when  $J_{n-1|n-1}$  is available in terms of the coefficients,  $b_{r\ell}$ . The computation task is then reduced to simply evaluating the coefficients for the updated density function, as shown.

The normalizing constant is omitted from the following sequence; the result is then correct to within a multiplicative constant, which is evaluated and inserted in the computer program.

The modal matrix of the covariance matrix of the prior cycle, designated  $Q$ , is normalized to make it an orthogonal matrix

$$Q' = Q^{-1} \quad (53)$$

We designated vectors in characteristic vector coordinates as  $\underline{v}$ , and in the position and velocity coordinates as  $\underline{x}$ . Then

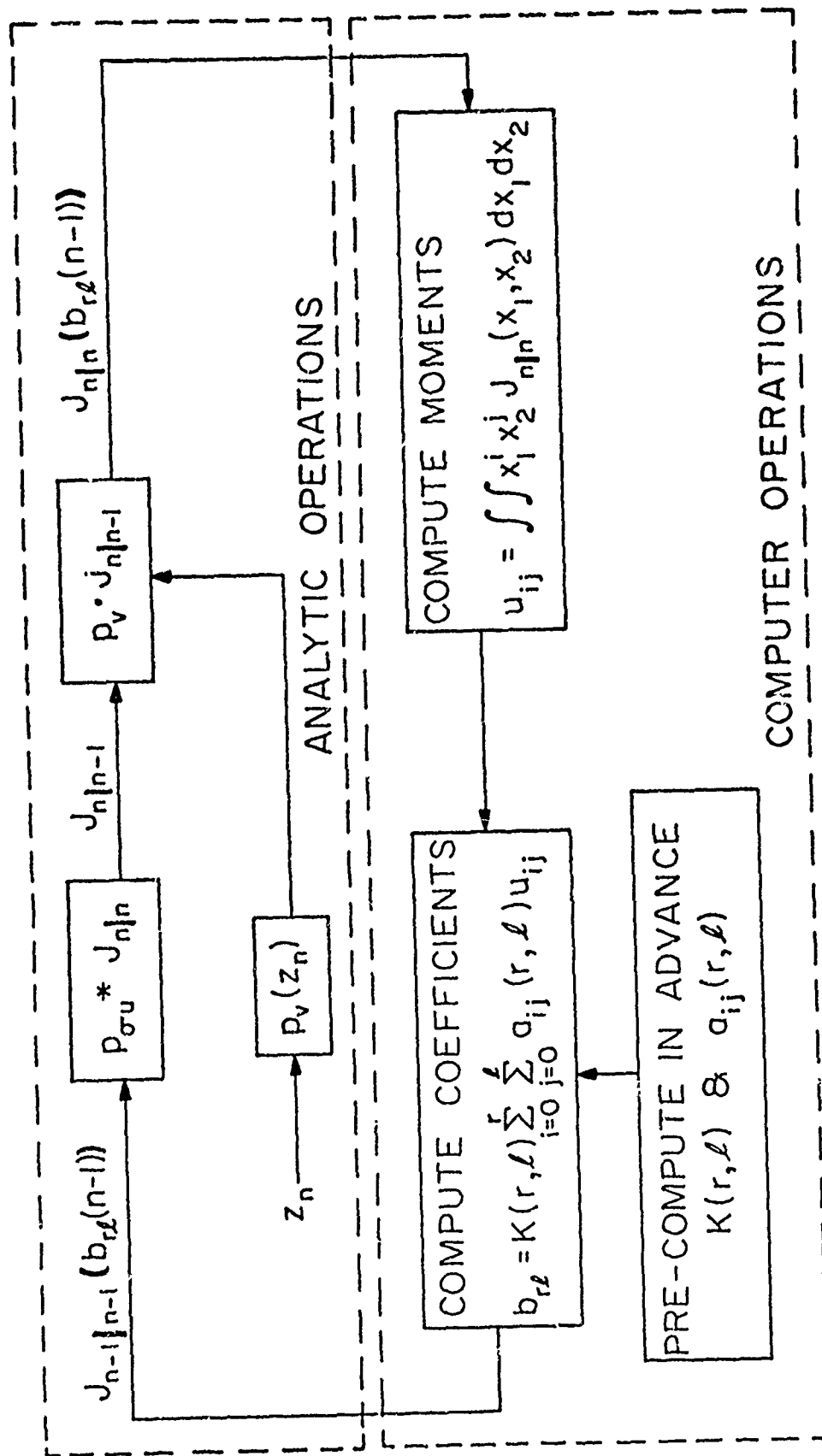


Fig. 1. Hermite Polynomial Bayes-Law Recursion

$$\underline{v} = T\underline{x} = Q^{-1}\underline{x} = Q'\underline{x}, \quad (54)$$

where

$$T = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix}$$

The polynomial expansion of the two-dimensional density is given by

$$J_{n|n} \begin{pmatrix} v_1(x) \\ v_2(x) \end{pmatrix} = \exp \left\{ -\frac{1}{2\lambda_1} (v_1 - \bar{v}_1)^2 - \frac{1}{2\lambda_2} (v_2 - \bar{v}_2)^2 \right\} \\ \sum_{r=0}^{m_1} \sum_{\ell=0}^{m_2} b_{r\ell} H_r \left[ \frac{1}{\sqrt{2\lambda_1}} (v_1 - \bar{v}_1) \right] H_\ell \left[ \frac{1}{\sqrt{2\lambda_2}} (v_2 - \bar{v}_2) \right] \quad (55)$$

In (55)  $\lambda_1, \lambda_2$  are the characteristic values of the covariance matrix, and  $\bar{v}_1, \bar{v}_2$  are the expected values.

The corresponding manipulations required to describe the prediction density  $J_{n+1|n}$  begin with

$$J_{n|n} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \exp \left\{ -\frac{1}{2\lambda_1} \left( t_{11} x_1 - t_{12} x_2 - \bar{v}_1 \right)^2 - \frac{1}{2\lambda_2} \left( t_{21} x_1 + t_{22} x_2 - \bar{v}_2 \right)^2 \right\} \\ \sum_{r=0}^{m_1} \sum_{\ell=0}^{m_2} b_{r\ell} H_r \left[ \frac{1}{\sqrt{2\lambda_1}} \left( t_{11} x_1 + t_{12} x_2 - \bar{v}_1 \right) \right] H_\ell \left[ \frac{1}{\sqrt{2\lambda_2}} \left( t_{21} x_1 + t_{22} x_2 - \bar{v}_2 \right) \right] \quad (56)$$

Next, we substitute  $y_1 - x_2 \Delta$  for  $x_1$  to prepare for the use of  $J_{n|n}$  in

(51). After some rearranging we obtain

$$J_{n+1|n} \begin{pmatrix} y_1 - x_2 \Delta \\ x_2 \end{pmatrix} = \sum_{r=0}^{m_1} \sum_{\ell=0}^{m_2} b_{r\ell} F_r(x_2) F_\ell(x_2) \quad , \quad (57)$$

where we have defined the terms

$$\begin{aligned} F_r(x_2) &\triangleq \exp \left\{ -\frac{1}{2\lambda_1 T_1^2} \left[ x_2 - \left( T_1 \bar{v}_1 - T_2 y_1 \right) \right]^2 \right\} \\ H_r &\left\{ \frac{\text{sign } T_1}{\sqrt{2\lambda_1 T_1^2}} \left[ x_2 - \left( T_1 \bar{v}_1 - T_2 y_1 \right) \right] \right\} \\ F_\ell(x_2) &\triangleq \exp \left\{ -\frac{1}{2\lambda_2 T_3^2} \left[ x_2 - \left( T_3 \bar{v}_2 - T_4 y_1 \right) \right]^2 \right\} \\ H_\ell &\left\{ \frac{\text{sign } T_3}{\sqrt{2\lambda_2 T_3^2}} \left[ x_2 - \left( T_3 \bar{v}_2 - T_4 y_1 \right) \right] \right\} \quad , \quad (58) \end{aligned}$$

with  $T_1, T_2, T_3$ , and  $T_4$  given by

$$T_1 \triangleq \frac{1}{t_{12} - t_{11} \Delta} = \text{sign}(T_1) \text{Abs}(T_1) \quad ,$$

$$T_2 \triangleq \frac{t_{11}}{t_{12} - t_{11} \Delta} \quad ,$$

$$T_3 \triangleq \frac{1}{t_{22} - t_{21} \Delta} = \text{sign}(T_3) \text{Abs}(T_3) \quad ,$$

and

$$T_4 \triangleq \frac{t_{21}}{t_{22} - t_{21} \Delta} \quad .$$

Now it happens that the prediction density (51) is a convolution with respect to  $y_2$ , so that we may combine (57) with (51) and let the asterisk denote convolution to obtain

$$\begin{aligned}
 J_{y_1}(y_2) &= \exp \left\{ -\frac{y_2^2}{2q\Delta} \right\} * J_{n|n} \left( \begin{array}{c} y_1 - y_2 \Delta \\ y_2 \end{array} \right) \\
 &= \exp \left\{ -\frac{y_2^2}{2q\Delta} \right\} * \sum_{r=0}^{m_1} \sum_{\ell=0}^{m_2} b_{r\ell} F_r(y_2) F_\ell(y_2). \quad (59)
 \end{aligned}$$

Finally, we are ready to take the Fourier transform of both sides of (59), giving

$$\begin{aligned}
 \mathcal{F}_{y_2} \left[ J_{n+1|n} \left( \begin{array}{c} y_1 \\ y_2 \end{array} \right) \right] &= \mathcal{F}_{y_2} \left[ \exp \left\{ -\frac{1}{2q\Delta} y_2^2 \right\} \right] \sum_{r=0}^m \sum_{\ell=0}^m b_{r\ell} \mathcal{F}_{y_2} \left[ F_r(y_2) \right] * \\
 &\quad \mathcal{F}_{y_2} \left[ F_\ell(y_2) \right] \quad (60)
 \end{aligned}$$

The algebraic details of evaluating the transforms are given by Hecht [9]. The result may be expressed as a function of the transform variable  $\omega$  as

$$\mathcal{F}_{y_2} \left[ J_{n+1} \left| n \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right. \right] = \sum_{r=0}^{m_1} \sum_{\ell=0}^{m_2} b_{r\ell} a_r c_\ell \sum_{k=0}^r \binom{r}{k} \exp \left\{ -i\omega \left( T_{11} \bar{v}_1 - T_{21} y_1 \right) \right\} \omega^{r-k} \exp \left\{ -\frac{1}{2} A \omega^2 \right\} * \\ \exp \left\{ -i\omega \left[ \left( 1 - \frac{1}{B} \right) \left( T_{11} \bar{v}_1 - T_{21} y_1 \right) + \frac{1}{B} \left( T_{32} \bar{v}_2 - T_{42} y_2 \right) \right] \right\} \left( 1 - \frac{1}{B} \right)^k \omega^{k+\ell} \left( \frac{1}{B} \right)^{\ell+1} \exp \left\{ -\frac{1}{2} \frac{C^2}{B^2} \omega^2 \right\},$$

(61)

where the following definitions have been used:

$$a_r \triangleq (-1)^r 2^{\frac{r}{2}} \left( \lambda_1 T_{11}^2 \right)^{\frac{r+1}{2}},$$

$$c_\ell \triangleq (-1)^\ell 2^{\frac{\ell}{2}} \left( \lambda_2 T_{33}^2 \right)^{\frac{\ell+1}{2}},$$

$$A \triangleq q\Delta + \lambda_1 t_1^2,$$

$$B \triangleq \frac{\lambda_1 t_1^2}{q\Delta + \lambda_1 t_1^2},$$

and

$$C^2 \triangleq \lambda_2 T_{33}^2 + Bq\Delta.$$

The quantity we desire, of course, is the inverse transform of (61), which we may obtain from (61) by taking advantage yielding finally [9]

$$\begin{aligned}
J_{n+1|n} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} &= \exp \left\{ -\frac{1}{2A} \left[ y_2 - T_1 \bar{v}_1 - T_2 y_1 \right]^2 \right\} \\
&\cdot \exp \left\{ -\frac{B^2}{2C^2} \left[ y_2 - \beta \left( T_1 \bar{v}_1 - T_2 y_1 \right) + \frac{1}{B} \left( t_4 y_1 - t_3 \bar{v}_2 \right) \right]^2 \right\} \\
&\cdot \sum_{r=0}^{m_1} \sum_{\ell=0}^{m_2} b_{r\ell} \left( \text{sign } T_1 \right)^r \left( \text{sign } T_3 \right)^\ell \frac{B^{\frac{r+1}{2}}}{B^{\frac{r+1}{2}}} \left( \frac{\lambda T_2^2 A}{\lambda T_2^2 q\Delta + \lambda T_1 T_2^2 + \lambda T_1^2 q\Delta} \right)^{\frac{\ell+1}{2}} \\
&\cdot \sum_{k=0}^r \binom{r}{k} (-1)^k \left[ \frac{(q\Delta)^2}{\lambda T_2^2 q\Delta + \lambda T_1 T_2^2 + \lambda q\Delta T_1^2} \right]^{k/2} \\
&\cdot H_{r-k} \left\{ \frac{1}{\sqrt{2A}} \left[ y_2 - \left( T_1 \bar{v}_1 - T_2 y_1 \right) \right] \right\} H_{k+\ell} \left\{ \frac{B}{\sqrt{2C}} \left[ y_2 + \beta \left( T_2 y_1 - T_1 \bar{v}_1 \right) + \frac{1}{B} \left( T_4 y_1 - T_3 \bar{v}_2 \right) \right] \right\},
\end{aligned}
\tag{62}$$

where we have used the additional definition

$$\beta^\Delta = \left( 1 - \frac{1}{B} \right) = 1 - \frac{q\Delta + \lambda T_1^2}{\lambda T_1^2} = - \frac{q\Delta}{\lambda T_1^2}.$$

$J_{n+1|n}$ , as given by (62), together with  $J_{n|n}$ , as given by (51), constitute the required pair of equations to update the conditional density for each sampling time.

To complete the cycle it is necessary to approximate the density function  $J_{n|n}$  with a new Hermite function as characterized by a new set of coefficients,  $b_{r\ell}$ .

In the previous Sections it was shown that under the proper conditions a function of two variables,  $f(x_1, x_2)$  can be approximated by a series expansion of Hermite polynomials of the form

$$f(x_1, x_2) \approx y(x_1, x_2) = e^{-\alpha_1^2 x_1^2} e^{-\alpha_2^2 x_2^2} \sum_{r=0}^{m_1} \sum_{\ell=0}^{m_2} b_{r\ell} H_r(\alpha_1 x_1) H_\ell(\alpha_2 x_2) \quad (63)$$

with the coefficients,  $b_{r\ell}$ , determined by

$$b_{r\ell} = \frac{\alpha_1 \alpha_2}{2^r r! 2^\ell \ell! \pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) H_r(\alpha_1 x_1) H_\ell(\alpha_2 x_2) dx_1 dx_2 \quad (64)$$

The approximation is the best in the sense that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\alpha_1^2 x_1^2} e^{-\alpha_2^2 x_2^2} \left[ f(x_1, x_2) - y(x_1, x_2) \right]^2 dx_1 dx_2 = \text{minimum} \quad (65)$$

The conditions for the approximation formulas to be valid are (Hildebrand [10]):

$$1. \quad e^{-\alpha_1^2 x_1^2} e^{-\alpha_2^2 x_2^2} \geq 0 \quad \text{in } (-\infty, \infty)$$

$$2. \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1^{k_{x_1}} x_2^{k_{x_2}} e^{-\alpha_1^2 x_1^2} e^{-\alpha_2^2 x_2^2} dx_1 dx_2 \text{ exist for all nonnegative integral}$$

values of  $k$ .

3. The integral in Equation (64) exists.

For the choice of  $\alpha_1^2 = \frac{1}{2\sigma_1^2}$  and  $\alpha_2^2 = \frac{1}{2\sigma_2^2}$ , with  $\sigma_1^2$  and  $\sigma_2^2$

the characteristic values of the covariance matrix, the first two conditions are obviously always true, and if  $f(x_1, x_2)$  is an exponentially decaying density function, condition 3 is true. In the developments of this section we associate  $f(x_1, x_2)$  with an approximating density function in the form of (51) with  $J_{n|n-1}$  given by (62). Since the coefficient function of  $J_{n|n-1}$  in (51) is bounded for all values of the arguments (and from reviewing the form of (62)), it is clear that condition 3 is true even for the approximating density function for any value of  $m$ .

Let  $f(x_1, x_2)$  be the density function we wish to approximate, and let  $m=0$ . Then

$$b_{00} = \frac{\alpha_1 \alpha_2}{\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 dx_2 = \frac{\alpha_1 \alpha_2}{\pi} \quad (66)$$

$$y(x_1, x_2) = \frac{\alpha_1 \alpha_2}{\pi} e^{-\alpha_1^2 x_1^2} e^{-\alpha_2^2 x_2^2} \quad (67)$$

Let  $\sigma_1^2$  and  $\sigma_2^2$  be the variances of  $f(x_1, x_2)$ , and let

$$\alpha_1^2 = \frac{1}{2\sigma_1^2}$$

$$\alpha_2^2 = \frac{1}{2\sigma_2^2}$$

Equation (67) is

$$y(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_1\sigma_2}} e^{-\frac{x_1^2}{2\sigma_1^2}} e^{-\frac{x_2^2}{2\sigma_2^2}} \quad (68)$$

We see that by taking only the zero'th term of the expansion we can get the best gaussian fit to the true density function, and also a verification on the form of the coefficient equation (64).

In addition to the above assumptions (except let  $m>0$ ), let  $\eta_1$  and  $\eta_2$  be the means of  $f(x_1, x_2)$ , and let  $\mu_{ij}$  be the  $i$ - $j$ <sup>th</sup> central moment.

Then

$$\mu_{ij} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 - \eta_1)^i (x_2 - \eta_2)^j f(x_1, x_2) dx_1 dx_2 \quad (69)$$

The product of the Hermite Polynomials when expanding about the expected value can be written

$$H_r[\alpha_1(x_1 - \eta_1)] H_\ell[\alpha_2(x_2 - \eta_2)] = \sum_{i=0}^r \sum_{j=0}^{\ell} r_{\ell} a_{ij} \alpha_1^i \alpha_2^j (x_1 - \eta_1)^i (x_2 - \eta_2)^j \quad (70)$$

with  $r_{\ell} a_{ij}$  a function of  $r$ ,  $\ell$ ,  $i$  and  $j$ .

Equation (64) is then rewritten as

$$\begin{aligned} b_{r\ell} &= \frac{\alpha_1 \alpha_2}{2^r r! 2^\ell \ell! \pi} \iint_{-\infty}^{\infty} f(x_1, x_2) \sum_{i=0}^r \sum_{j=0}^{\ell} r_{\ell} a_{ij} \alpha_1^i \alpha_2^j (x_1 - \eta_1)^i (x_2 - \eta_2)^j dx_1 dx_2 \\ &= \frac{\alpha_1 \alpha_2}{2^r r! 2^\ell \ell! \pi} \sum_{i=0}^r \sum_{j=0}^{\ell} r_{\ell} a_{ij} \alpha_1^i \alpha_2^j \mu_{ij} \end{aligned} \quad (71)$$

Comparing (63) with (60), we see the approximating polynomial expansion was accomplished by letting

$$\alpha_1^2 = \frac{1}{2\lambda_1}, \quad \alpha_2^2 = \frac{1}{2\lambda_2}, \quad \eta_1 = \bar{v}_1, \quad \eta_2 = \bar{v}_2,$$

and using (71) to compute the coefficients,  $b_{r\ell}$ . The coefficients,  $r\ell^{a_{ij}}$ , in (70) and (71) are functions of the coefficients of the Hermite Polynomials, and can be computed one time in advance and stored. The formula for generating the Hermite coefficients is given in the previous section. Designating  $c_{mn}$  as the coefficient of  $x^n$  in  $H_m(x)$ ,  $r\ell^{a_{ij}}$  in Equation (70) is determined for each  $r$  and for each  $\ell$  from

$$r\ell^{a_{ij}} = c_{ri}c_{\ell j}$$

### 5. Applying the Hermite Expansion

The general formulas of the previous section are specialized here for the case when the system equations are those of the phase demodulator (see Chapter VII). This section describes the techniques that were used to mechanize the given equations. In particular, the conditional means  $\bar{v}_i$ , the characteristic values  $\lambda_i$ , and vectors,  $t_{ij}$ , of the covariance matrix, and the coefficients of the Hermite approximation,  $b_{r\ell}$  are described in detail for this specific example.

To perform the integrations to obtain the means and central moments (to compute the quantities in the above paragraph) Equations (51) and (62) are combined into one equation, omitting temporarily the normalizing constant.

$$J_{n|n} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = S(y_1 y_2) J_{n|n-1} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad (73)$$

where

$$S(y_1 y_2) = \exp \left\{ -\frac{\Delta}{2r} \left[ (z_1 - \cos y_1)^2 + (z_2 - \sin y_1)^2 \right] \right\} \quad (74)$$

and  $J_{n|n-1}$ , given by (62), consists of two parts, an exponential function and a polynomial function of the arguments  $y_1$  and  $y_2$ .

The exponent of the exponential part was a quadratic form; i.e., of the form  $\{-a ||x-\bar{x}||_A^2\}$ . To perform integrations of (73) and Equation (73) times  $y_1^i y_2^j$ , the combined exponential function is put into the following form in order to use the Gauss-Hermite formula:

$$J_{n|n} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \exp \left\{ -\frac{1}{2} \left[ \frac{(y_1 - m_1)^2}{\sigma_1^2 (1 - \rho^2)} - \frac{2\rho (y_1 - m_1)(y_2 - m_2)}{\sigma_1 \sigma_2 (1 - \rho^2)} + \frac{(y_2 - m_2)^2}{\sigma_2^2 (1 - \rho^2)} \right] \right\} F(y_1 y_2) \quad (75)$$

with the exponent of  $S(y_1 y_2)$  of (73) (as given by (51)), linearized, and the linear terms combined with the linear exponential terms of  $J_{n|n-1}(y_1 y_2)$ . The error between the linear and nonlinear parts of the exponent were combined with the polynomial part of  $J_{n|n-1}$  to form  $F(y_1 y_2)$ . This technique was suggested in Bucy, Geesey, and Senne [5].

The terms of (75) are

$$m_1 \triangleq E(y_1)$$

$$m_2 \triangleq E(y_2)$$

$$\sigma_1^2 \triangleq E[(y_1 - m_1)^2]$$

$$\sigma_2^2 \triangleq E[(y_2 - m_2)^2]$$

$$\rho = \text{correlation coefficient} = \frac{E[(y_1 - m_1)(y_2 - m_2)]}{\sigma_1 \sigma_2}$$

We operate on the exponential part of (74) as follows.

$$\begin{aligned} & - \frac{\Delta}{2r} \left[ (z_1 - \cos y_1)^2 + (z_2 - \sin y_1)^2 \right] \\ = & - \frac{\Delta}{2r} (z_1 - \cos y_1 - \overline{\cos y_1} + \overline{\cos y_1}) - \frac{\Delta}{2r} (z_2 - \sin y_1 \\ & - \overline{\sin y_1} + \overline{\sin y_1})^2 \end{aligned} \quad (76)$$

with

$$\overline{\cos y_1} = \cos y_1^* + y_1^* \sin y_1^* - y_1 \sin y_1^*$$

$$\overline{\sin y_1} = \sin y_1^* - y_1^* \cos y_1^* + y_1 \cos y_1^*$$

$$y_1^* = E[y_1(n) | z(n-1), \dots, z(0)]$$

$$\begin{aligned}
& \exp \left\{ - \frac{\Delta}{2r} \left[ (z_1 - \cos y_1)^2 + (z_2 - \sin y_1)^2 \right] \right\} \\
& = \exp \left\{ - \frac{\Delta}{2r} (z_1 - \overline{\cos y_1})^2 - \frac{\Delta}{2r} (z_2 - \overline{\sin y_1})^2 \right\} \\
& \quad \exp \left\{ - \frac{\Delta}{2r} \left[ (\overline{\cos y_1} - \cos y_1)^2 + 2(z_1 - \overline{\cos y_1})(\overline{\cos y_1} - \cos y_1) \right. \right. \\
& \quad \left. \left. + (\overline{\sin y_1} - \sin y_1)^2 + 2(z_2 - \overline{\sin y_1})(\overline{\sin y_1} - \sin y_1) \right] \right\}
\end{aligned}
\tag{77}$$

The first line on the right of the equal sign is the linearized exponent which is combined with the exponential terms of Equation (62). The second and third lines are reworked to give

$$\begin{aligned}
& \exp \left\{ \frac{\Delta}{2r} (y_1 - y_1^*)^2 + \frac{z_1 \Delta}{r} \cos y_1 + \frac{z_2 \Delta}{r} \sin y_1 + \frac{z_1 \Delta}{r} (y_1 - y_1^*) \sin y_1^* \right. \\
& \quad \left. - \frac{z_1 \Delta}{r} \cos y_1^* - \frac{z_2 \Delta}{r} \sin y_1^* + \frac{z_2 \Delta}{r} (y_1^* - y_1) \cos y_1^* \right\}
\end{aligned}
\tag{78}$$

as the exponential part of the function to be evaluated for the integration. That is  $F(y_1 y_2) = [\text{exponent of (78)}]$ . (polynomial part of  $J_n|_{n-1}$ ).

The required integrals are of the following form

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_1^i y_2^j J_n \left| n \left( \frac{y_1}{y_2} \right) \right| dy_1 dy_2 \quad (79)$$

Thus we compute

$$J_m \left| m \left( \frac{y_1}{y_2} \right) \right| = \exp \left\{ -\frac{1}{2} \left[ \frac{(y_1 - m_1)^2}{\sigma_1^2 (1 - \rho^2)} - \frac{2\rho (y_1 - m_1)(y_2 - m_2)}{\sigma_1 \sigma_2 (1 - \rho^2)} + \frac{(y_2 - m_2)^2}{\sigma_2^2 (1 - \rho^2)} \right] \right\} F(y_1 y_2) ,$$

$$\begin{aligned} F(y_1 y_2) = \exp \left\{ -\frac{\Delta}{2r} (y_1 - y_1^*)^2 + \frac{z_1 \Delta}{r} \cos y_1 + \frac{z_2 \Delta}{r} \sin y_1 \right. \\ \left. + \frac{z_1 \Delta}{r} (y_1 y_1^*) \sin y_1^* - \frac{z_1 \Delta}{r} \cos y_1^* - \frac{z_2 \Delta}{r} \sin y_1^* \right. \\ \left. + \frac{z_2 \Delta}{r} (y_1^* - y_1) \cos y_1^* \right\} \end{aligned}$$

$$\sum_{\eta=0}^m \sum_{\ell=0}^m b_{r\ell} (\text{sign } T_1)^r (\text{sign } T_3)^\ell (B)^{\frac{r+\ell}{2}} \left( \frac{\lambda_2 T_3^2 A}{\lambda_2 T_3^2 q + \lambda_1 \lambda_2 T_1^2 T_3^2 + \lambda_1 T_1^2 q} \right)^{\frac{\ell+1}{2}}$$

$$\sum_{k=0}^r \binom{r}{k} (-1)^k \left[ \frac{(q\Delta)^2}{\lambda_2 T_3^2 q + \lambda_1 \lambda_2 T_1^2 T_3^2 + \lambda_1 q \pi} \right]^{k/2} .$$

$$H_{r-k} \left\{ \frac{1}{\sqrt{2A}} \left[ y_2 - (T_1 \bar{v}_1 - T_2 y_1) \right] \right\} H_{k+\ell} \left\{ \frac{1}{\sqrt{2A}} \left[ 2^{-\beta} (T_2 y_1 - T_1 \bar{v}_1) + \frac{1}{B} (T_4 y_1 - T_3 \bar{v}_2) \right] \right\} , \quad (80)$$

and

$$\begin{aligned}
 & \exp \left\{ -\frac{1}{2} \left[ \frac{(y_1 - m_1)^2}{\sigma_1^2(1-\rho^2)} - \frac{2\rho(y_1 - m_1)(y_2 - m_2)}{\sigma_1\sigma_2(1-\rho^2)} + \frac{(y_2 - m_2)^2}{\sigma_2^2(1-\rho^2)} \right] \right\} \\
 &= \exp \left\{ -\frac{\Delta}{2r}(z_1 - \overline{\cos y_1}) - \frac{\Delta}{2r}(z_2 - \overline{\sin y_1}) - \frac{1}{2} \left[ y_2 - (T_1 \bar{v}_1 - T_2 y_1) \right]^2 \right. \\
 &\quad \left. - \frac{B^2}{2C^2} \left[ y_2 - \beta(T_1 \bar{v}_1 - T_2 y_1) + \frac{1}{B}(T_4 y_1 - T_3 \bar{v}_2) \right]^2 \right\} \quad (81)
 \end{aligned}$$

Let the following abbreviations be made in (81):

$$\begin{aligned}
 v_4 &= T_1 \bar{v}_1 \\
 v_5 &= -\beta T_1 \bar{v}_1 - \frac{T_3}{B} \bar{v}_2 \\
 v_6 &= \frac{T_4}{B} + \beta T_2
 \end{aligned}$$

and equate the coefficients of  $y_1^2$ ,  $y_2^2$ ,  $y_1 y_2$ ,  $y_1$  and  $y_2$  in (26), thus solving for  $\sigma_1^2$ ,  $\sigma_2^2$ ,  $\rho$ ,  $m_1$ , and  $m_2$ .

$$\frac{1}{\sigma_1^2(1-\rho^2)} = c_3 = \frac{T_2^2}{A} + \frac{B^2}{C^2} v_6^2 + \frac{1}{R}, \quad (82)$$

$$\frac{1}{\sigma_2^2(1-\rho^2)} = c_2 = \frac{1}{A} + \frac{B^2}{C^2}, \quad (83)$$

$$\frac{-\rho}{\sigma_1\sigma_2(1-\rho^2)} = c_1 = \frac{T_2}{A} + \frac{B^2}{C^2} v_6, \quad (84)$$

$$\frac{-m_1}{\sigma_1^2(1-\rho^2)} + \frac{\rho m_2}{\sigma_1 \sigma_2(1-\rho^2)} = c_4$$

$$= \frac{A}{r} (z_1 \sin y_1^* - z_2 \cos y_1^* - y_1^*) - \frac{T_2 v_{14}}{A} + \frac{B^2}{C^2} v_5 v_6 \quad (85)$$

$$- \frac{m_2}{\sigma_2^2(1-\rho^2)} + \frac{\rho m_1}{\sigma_1 \sigma_2(1-\rho^2)} = c_5 = - \frac{v_{14}}{A} + \frac{B^2}{C^2} v_5 \quad , \quad (86)$$

from which we determine that

$$\rho = - \frac{c_1}{\sqrt{c_2} \sqrt{c_3}} \quad , \quad (87)$$

$$\sigma_1^2 = \frac{1}{c_3(1-\rho^2)} \quad , \quad (88)$$

$$\sigma_2^2 = \frac{1}{c_2(1-\rho^2)} \quad , \quad (89)$$

$$m_1 = - \frac{1}{(1-\rho^2)} \left( \frac{c_4}{c_3} + \frac{\rho c_5}{\sqrt{c_2} \sqrt{c_3}} \right) \quad , \quad (90)$$

and

$$m_2 = - \frac{1}{(1-\rho^2)} \left( \frac{c_5}{c_2} + \frac{\rho c_4}{\sqrt{c_2} \sqrt{c_3}} \right) \quad . \quad (91)$$

Using the Gauss-Hermite formula

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp \left\{ -x_1^2 - x_2^2 \right\} G(x_1, x_2) dx_1 dx_2 \approx \sum_{i=1}^m \sum_{j=1}^m w_{1i} w_{2j} G(x_{1i}, x_{2j}) \quad (92)$$

the mechanics of the integrations are as follows:

1. The eigenvalues and eigenvectors of the covariance matrix of the linearized equations,

$$\begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix}$$

are evaluated. Figure 2 illustrates the various axes directions described here.

2. The formula (92), is used with the grid points scaled proportional to  $\sqrt{2\lambda_1}$  and  $\sqrt{2\lambda_2}$  and the directions along the corresponding eigenvectors are referred to as the rotated coordinates.

That is,

$$x_{1i} = \sqrt{2\lambda_1} t_i$$

$$x_{2i} = \sqrt{2\lambda_2} t_i$$

$$i = 1, n$$

$n$  = number of points for integration in each dimension with  $t_i$  the abscissa value given in the numerical integration tables.

$(m_1, m_2) = (E(y_1), E(y_2))$  FOR LINEARIZED OBSERVATION

$x_1, x_2$  = EIGENVECTORS FOR " "

= ROTATED COORDINATES

$(\bar{m}_1, \bar{m}_2) = (E(y_1), E(y_2))$  FOR NONLINEARIZED OBSERVATION

$e_1, e_2$  = EIGENVECTORS FOR " "

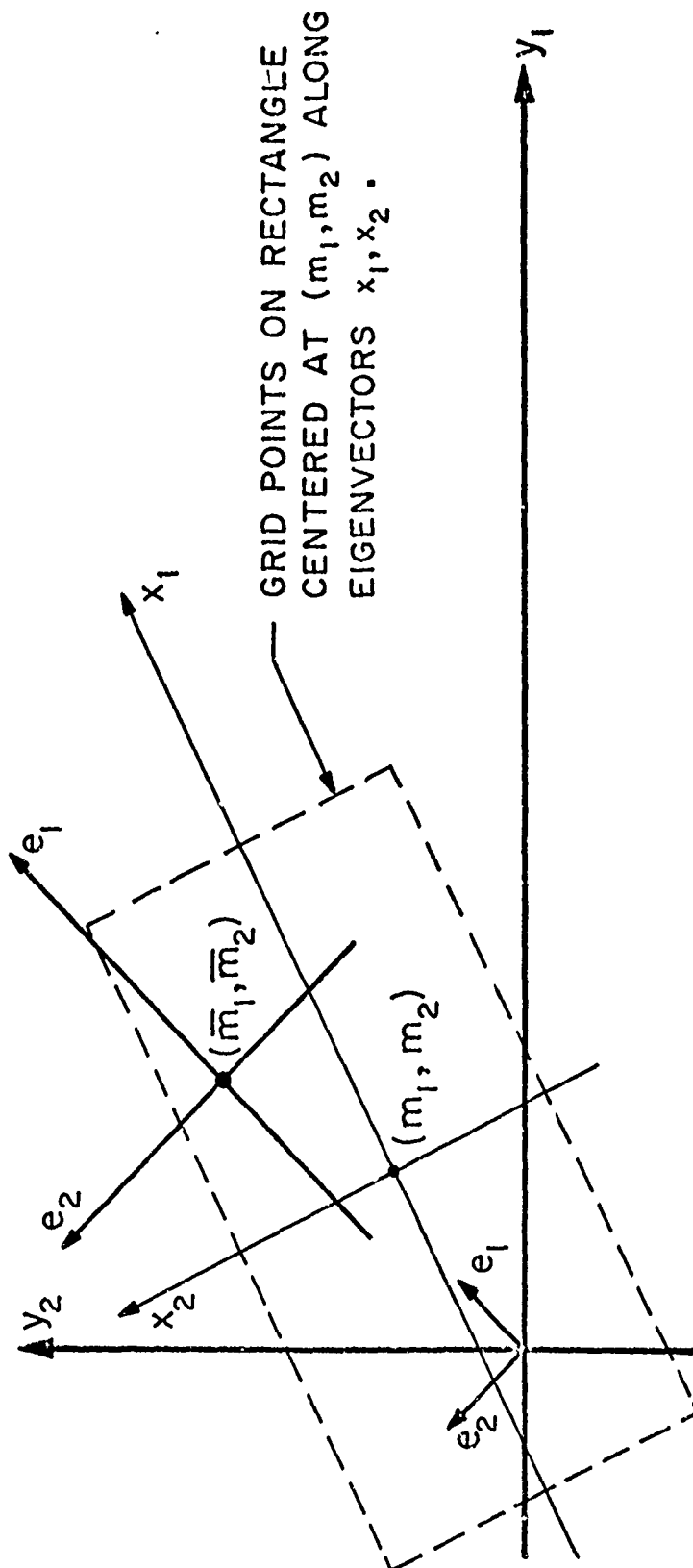


Fig. 2. Coordinate Systems for Hermite Expansion

3. The points  $x_{1i}$  and  $x_{2i}$  are rotated back to the coordinate axis using the transformation matrix formed by the eigenvectors. The points are then shifted to be centered about  $m_1$  and  $m_2$ .
4. The function (80) is evaluated at each of the  $n^2$  points,  $y_{1i}$ ,  $y_{2i}$ .
5. Moments of the required degree,  $k$  and  $\ell$ , were found from

$$I_{k\ell} = \frac{1}{I_{00}} \sum_{i=1}^m \sum_{j=1}^m y_{1i}^k y_{2j}^\ell w_{1i} w_{2j} F(y_{1i} y_{2j}), \quad (93)$$

where

$$I_{00} = \sum_{i=1}^m \sum_{j=1}^m w_{1i} w_{2j} F(y_{1i} y_{2j})$$

6. The means and central moments (in the rotated coordinates) are determined according to the formula (94) given below.
7. The eigenvalues and eigenvectors of the covariance matrix formed by the second central moments, in the rotated coordinates, are evaluated.
8. The central moments,  $\mu_{ij}$ , required for the Hermite expansion in the directions of the eigenvectors determined in Step 7, using the results of Step 6, are evaluated, using the formula (94) given below.
9. The equations of Section I are used to determine a new set of coefficients,  $b_{r\ell}$ , to characterize the new density function.

To perform the computations of the above steps, two relations are needed: 1) a formula to convert moments to central moments, and 2) a formula to compute central moments about a rotated set of coordinates. The first formula is given as follows:

Given the two-dimensional moments

$$\begin{aligned}
 & E (x_1^i x_2^i) \quad \text{and means} \quad \eta_1, \eta_2 \quad ; \\
 & E \left[ (x_1 - \eta_1)^m (x_2 - \eta_2)^m \right] \\
 & = E \left[ \sum_{r=0}^m \binom{m}{r} x_1^{m-r} (-\eta_1)^r \sum_{s=0}^m \binom{m}{s} x_2^{m-s} (-\eta_2)^s \right] \\
 & = \sum_{n=0}^m \sum_{s=0}^m \binom{m}{r} \binom{n}{s} (-\eta_1)^r (-\eta_2)^s E \left[ (x_1^{m-r}) (x_2^{m-s}) \right] \quad (94)
 \end{aligned}$$

The second formula was found as follows:

Given central moments in x-axes and a transformation matrix A  
with

$$\begin{aligned}
 \underline{y} &= A \underline{x} \\
 A &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}
 \end{aligned}$$

we have

$$\begin{aligned}
 E [y_1^m y_2^n] &= E \left[ (a_{11}x_1 + a_{12}x_2)^m (a_{21}x_1 + a_{22}x_2)^n \right] \\
 &= E \left[ \sum_{r=0}^m \binom{m}{r} (a_{11}x_1)^{m-r} (a_{12}x_2)^r \sum_{s=0}^n \binom{n}{s} (a_{21}x_1)^{n-s} (a_{22}x_2)^s \right] \\
 &= \sum_{r=0}^n \sum_{s=0}^m \binom{n}{r} \binom{m}{s} a_{11}^{m-r} a_{12}^r a_{21}^{n-s} a_{22}^s E [x_1^{m+n-r-s} x_2^{r+s}] \quad (95)
 \end{aligned}$$

Upon computing the coefficients,  $b_{rl}$ , for the Hermite expansion, the following observations are made:

1. The coefficient  $b_{00}$  is, from (66),

$$b_{00} = \frac{1}{\pi} (2\lambda_1)^{-1/2} (2\lambda_2)^{-1/2}$$

= normalizing constant for a gaussian density with variances  $\lambda_1, \lambda_2$ .

2.  $b_{10} = b_{01} = 0$ , from (71).
3.  $b_{11} = 0$ , from (71). When expanding in eigenvector coordinates  $\mu_{11} = 0$ .
4.  $b_{20} = b_{02} = 0$ , from (71). The Hermite polynomials are (Section 2)

$$H_0(x) = 1$$

$$H_1(x) = 2x$$

$$H_2(x) = 4x^2 - 2$$

To compute  $b_{20}$ , for example, using (71),

$$b_{20} = b_{00} \frac{1}{2^2 2!} \left[ 2_0 a_{00} + 2_0 a_{10} (2\lambda_1)^{-1/2} \mu_{10} + 2_0 a_{20} (2\lambda_1)^{-1} \mu_{20} \right]$$

where from (72)

$${}_{20}a_{00} = -2$$

$${}_{20}a_{10} = 0$$

$${}_{20}a_{20} = 4$$

$$b_{20} = b_{00} \frac{1}{2^2 2!} \left[ -2 + 2\lambda_1^{-1} \mu_{20} \right] = 0$$

since  $\mu_{20} = \lambda_1$ ,

5. Noting  $b_{00}$  is only a normalizing constant, the lowest coefficients,  $b_{r\ell}$ , which contains information beyond the gaussian fit are  $b_{30}$ ,  $b_{21}$ ,  $b_{12}$ , and  $b_{03}$ . The reason for this is intuitively clear.

a)  $b_{10}$  and  $b_{01}$  account for the expected value of the approximating gaussian function being different than the actual function.

b)  $b_{11}$  accounts for the lowest order cross term of the approximating gaussian function being different than zero.

c)  $b_{20}$  and  $b_{02}$  account for the variances of the approximating gaussian function being different than the actual function.

6. The Hermite expansion, as given in Section 2, is such that addition of higher terms of a truncated series does not affect lower order coefficients (Wiener [14]). For example, in (63), if  $m$  were changed to  $m+1$ ,  $b_{r\ell} (r \leq m, \ell \leq m)$  would not change.

When developing the computer program, experiments were made to determine the number of terms to carry in the series expansion, i.e., the value of  $m$  in (63). The value was related to the extent of the nonlinearities. For the phase-lock problem (Chapter VII) the non-

linearity was approximately proportional to the parameter  $p_{11}(0) = \sqrt{2} r^{3/4} q^{1/4}$ . For  $p_{11}(0)^{1/2} < 0.1$  radian the linear and nonlinear programs showed insignificant differences. For  $.1 < p_{11}(0)^{1/2} < .55$  radian, tests were made for  $m$  ranging to 9. There was little difference in the sequence of estimates between  $m=5$  and  $m=9$ . Monte Carlo tests were therefore made with  $m=5$ .

It was further noted in the experiments that higher order cross terms could be neglected for equivalent accuracy. In (63), instead of

$$r = (0,5)$$

$$l = (0,5)$$

it was sufficient to let  $r, l =$  nonnegative integers such that  $r+l=50$ .

This was reasonable in view of (64), where it may be noted  $b_{r,l}$  is normalized by a factor  $\frac{1}{2^r r! 2^l l!}$ . The coefficients  $b_{50}$  or  $b_{05}$ , therefore, would carry 3840 times more weight than the coefficient  $b_{55}$ , if it were used.

One additional alteration was made as a result of the experiments, to enable the program to perform satisfactorily. Occasionally, it was observed, the density function would take on small negative values. To normalize the density function a multiplicative factor was used to make the integral of the function equal to one. This factor, of course, multiplied the negative values as well as the positive values, which created problems. To avoid the problem, the density function was set equal to zero whenever a negative value was computed. This meant that the Hermite approximation was no longer the best fit in the sense of the previous Sections. However, the negative values that were neglected were always small, never being greater than about  $10^{-4}$  times the largest positive value of the density function.

### C. The Point-Mass Approximation

The method of representing density functions by point masses on a floating grid was originated by Bucy [ 2 ], and refined by Bucy and Senne [ 6 ]. We abstract a portion of the latter reference here for an introduction of the general method. To begin we let  $J_n(\underline{y})$  be the one-step predictor density and consider the  $(2M+1)^d$  points represented by the expression

$$J_n(\underline{y}) = \sum_{i_1, i_2, \dots, i_d=1}^{2M+1} J_n[g_n(i_1, \dots, i_d)] \delta[\underline{y} - g_n(i_1, \dots, i_d)] \quad (96)$$

where  $\delta$  is the Dirac delta function of  $d$  arguments.

Define a vector  $J_n \in R^{(2M+1)^d}$  as

$$\underline{J}_n \triangleq \left\{ J_n[g_n(1, 1, \dots, 1)], \dots, J_n[g_n(2M+1, \dots, 2M+1)] \right\} \quad (97)$$

and the grid map

$$\underline{g}_n : K^d \rightarrow R^d \quad K = \{1, \dots, 2M+1\}.$$

With this approximation the state becomes a pair  $(\underline{J}_n, \underline{g}_n(\cdot))$ , the first a  $(2M+1)^d$  vector, and the second a function specified by a  $(2M+1)^d$  vector of its ordered images in  $R^d$ . In other words the effective state dimension becomes  $(d+1)(2M+1)^d$ , since the map  $\underline{g}_n$

is determined by a  $dx(2M+1)^d$  table or matrix. Now substituting (96) in Bayes law we arrive at the state update equation.

$$C_{n-n+1}^J(i_1, \dots, i_d) = \sum_{j_1, \dots, j_d=1}^{2M+1} T_n[g_{n+1}(i_1, \dots, i_d), g_n(j_1, \dots, j_d)] J_n(j_1, \dots, j_d) \quad (98)$$

where

$$T_n(g_{n+1}, g_n) \triangleq \exp \left[ \langle z_n, h_n(z_n) \rangle R^{-1}(n) - 1/2 \| h_n(g_n) \|^2_{R^{-1}(n)} \right] N \left[ g_{n+1} - \phi_{n+1}(g_n), A_{n+1}(g_{n+1}) \right]$$

with

$$N(\underline{a}, B) \triangleq \frac{(\det B)^{-1/2}}{(2\pi)^{d/2}} \exp \left\{ -1/2 \| \underline{a} \|^2_{B^{-1}} \right\},$$

and  $C_n$  is chosen so that the total mass of  $J_{n+1}$  is one.

Now (98) can be sequentially computed by a digital computer, once the gridding is determined. However, we must evaluate  $(2M+1)^{2d}$  matrix elements and multiply a  $(2M+1)^d$  vector with the matrix. In order to select the gridding to make our approximation most effective we will center the grid for  $I_n(y)$  at our best estimate of  $x_n$  given  $z_{n-2}, \dots, z_0$  [i.e.  $\hat{x}(n|n-2)$ ] since the gridding for  $J_n$  must be given before  $J_n$  is computed and hence only  $J_{n-1}$  is known. Similarly, the mesh size

and directions are given in terms of the eigenvalues and eigenvectors of the conditional error covariance  $\Sigma(n|n-2)$  of  $\underline{x}_n$  given  $\underline{z}_{n-2}, \dots, \underline{z}_0$ . Explicitly these parameters are given by

$$\begin{aligned} \hat{\underline{x}}(n|n-2) &= \int d\underline{p} \int [\phi_n(\underline{\zeta}) + \Gamma_n(\underline{\zeta})\underline{\lambda}] J_{n-1}(\underline{\zeta}) \underline{\lambda} d\underline{\zeta} d\underline{\lambda} \\ &= \int d\underline{\zeta} \int \phi_n(\underline{\zeta}) J_{n-1}(\underline{\zeta}) d\underline{\zeta} \\ &= \sum_{i_1, \dots, i_d=1}^{2M+1} \phi_n[g_{n-1}(i_1, \dots, i_d)] J_{n-1}(i_1, \dots, i_d), \end{aligned} \quad (99)$$

where  $\{\lambda_n(\cdot)\}$  are the densities of the  $u(n)$  sequence.

$$\begin{aligned} \Sigma(n|n-2) &= \hat{\underline{x}}(n|n-2) \hat{\underline{x}}^T(n|n-2) \\ &= \int d\underline{\zeta} \int [\phi_n(\underline{\zeta}) \phi_n^T(\underline{\zeta}) + A_n(\underline{\zeta})] J_{n-1}(\underline{\zeta}) d\underline{\zeta} \\ &= \sum_{i_1, \dots, i_d=1}^{2M+1} \left\{ \phi_n[g_{n-1}(i_1, \dots, i_d)] \phi_n^T[g_{n-1}(i_1, \dots, i_d)] \right. \\ &\quad \left. + A_n[g_{n-1}(i_1, \dots, i_d)] \right\} J_{n-1}(i_1, \dots, i_d) \end{aligned} \quad (100)$$

Let  $\lambda_n^1, \dots, \lambda_n^d, e_n^1, \dots, e_n^d$  be the eigenvalues and eigenvectors of 67  
 $\Sigma(n|n-2)$ . Then define  $g_n$  by

$$g_n(i_1, \dots, i_d) = [e_n^1, \dots, e_n^d] \begin{bmatrix} 1 \\ K_n \\ \vdots \\ K_n^d \end{bmatrix} + \hat{x}(n|n-2), \quad (101)$$

where

$$K_n^k = n_{\sigma} (\lambda_n^k)^{1/2} \left( \frac{i_k - 1}{M} - 1 \right),$$

and  $n_{\sigma}$  is a constant parameter.

We call the grid  $g_n$  a floating grid. The floating grid is centered at the best available estimate of the current mean and rotated from the state coordinate frame into the principle axes of the best available estimate of the error ellipsoid.

In Bucy and Senne [6] there is a general discussion of the computational problems associated with the above scheme, and some general (problem independent) method for reducing the computational burden. It turns out, however, that the most impressive simplifications generally take place as a result of peculiarities of the application at hand. For example, the problem of Chapter VII in this report illustrates how one takes advantage of a singular A matrix (i.e. fewer driving noises than states). In Appendix C of Chapter VII we will pursue that example closely to determine the implications of the simplification on the point-mass calculations.

#### D. Non-Orthogonal Series ~ Gaussian Sums

In his dissertation Lo [12] observed that nonlinear filtering for linear systems with non-gaussian a priori densities can be achieved to any accuracy desired by approximating the a priori density with a suitable sum of weighted gaussians, and initializing a bank of linear Kalman-Bucy filters from the terms of the gaussian sum. Lo further showed that certain nonlinear systems with measurement functions possessing finite-dimensional sensor orbits (see Bucy and Joseph [4]) can be transformed by change of coordinates into linear systems with non-gaussian a priori densities. Thus optimal nonlinear filtering (at least in the sensor-orbit coordinates) is achievable for such problems. Unfortunately, the relationship between the optimal sensor-orbit estimate and the optimal estimate in the original coordinates is not simple, so that if one needs optimal nonlinear estimates in the original coordinates it is frequently advisable to retain the original coordinates and use Bayes Law, as discussed in the previous sections. The question arises, though, whether there is a suitable generalization of the bank of linear filters used by Lo which will apply in the original coordinates to nonlinear problems. Such an expansion would be equivalent to representing the a posteriori density by a nonorthogonal series of gaussians.

One advantage of a gaussian sum is that it must be everywhere positive, so that the truncated series will not have the unfortunate characteristic of oscillating negative/positive instabilities. On the other hand, since the series is nonorthogonal (on  $L_2$ ) it happens that the determination of the minimum-norm (in  $L_2$ ) gaussian sum approximation

requires inverting a large matrix of inner products. Center [6] has described the generalized least-squares process and has illustrated how the selection of coefficients for the representation amounts to solving an unconstrained minimum norm problem in  $L_2$  (i.e., an orthogonal projection). Thus, the optimal choice of coefficients for a non-orthogonal basis involves a substantial amount of computation.

Recognizing this computational burden, Alspach [1] proposed a sub-optimal scheme for dropping gaussians until a suitable fit (called a "theorem fit") is obtained. He discusses several special cases, such as the tracking problem in Chapter VI, where simplifications arise from symmetry of the density functions. Once the appropriate fit is obtained, however, there remains the problem of implementing Bayes Law. For this Alspach proposes local linearizations. He expands both the driving noise density and the measurement noise density in suitable gaussian sums, then essentially considers the pairs of points which occur as a result of Bayes Law, linearizes about these points and uses partial realizations of linearized Kalman-Bucy filters for each of the terms. The result is an increase in the number of gaussians during the Bayes Law computation, so that he must then reduce the resulting number of terms by discarding those with insignificant weight.

The above procedure may be summarized as shown in Table 1 (adapted from Alspach [1], pp. 101-110).

Table 1. Outline of a Gaussian-Sum  
Recursion Procedure [1]

A. Form  $p_{x_k}(\xi_k | z_{k-1}, \dots, z_0)$  as a gaussian sum approximation in the form

$$p_{x_k}(\xi_k | z_{k-1}, \dots, z_0) = \sum_{i=1}^{\eta'_k} \alpha'_{k_i} N(\xi_k - \underline{a}_{k_i}, P'_{k_i}), \quad (102)$$

where  $P'_{k_i}$  is constrained to have the property that for a preassigned  $s > 1$ ,

$$s^2 P'_{k_i} < E \text{ for all } i \in [1, \eta'_k],$$

so that  $E - s^2 P'_{k_i}$  is positive definite for all  $i$ .

B. Linearize  $h_k(x_k)$  about the mean of each term or each  $\underline{a}_{k_i}$ , forming

$$h_{k_i}(x_k) = h_k(\underline{a}_{k_i}) + H_k(\underline{a}_{k_i})(x_k - \underline{a}_{k_i})$$

Then Form  $p_{x_k}(\xi_k | z_k, \dots, z_0)$  as a gaussian sum as

$$p_{x_k}(\xi_k | z_k, \dots, z_0) = \sum_{j=1}^{\eta_k} \alpha_{k_j} N(\xi_k - \underline{\mu}_{k_j}, P_{k_j}), \quad (103)$$

where

$$j = 1, 2, \dots, \eta_k \quad \eta_k = \eta'_k,$$

$$\underline{\mu}_{k_j} = \underline{a}_{k_j} + K_{k_j} [z_k - h_k(\underline{a}_{k_j})],$$

$$P_{k_j} = P'_{k_j} - K_{k_j} H_k(\underline{a}_{k_j}) P'_{k_j},$$

and 
$$\alpha_{k_j} = \frac{\alpha'_{k_j} N(\underline{z}_k - \underline{H}_k(\underline{a}_{k_j}), \underline{H}_k(\underline{a}_{k_j}) P'_{k_j} \underline{H}_k(\underline{a}_{k_j})^T + R_k)}{\sum_{j=1}^{\xi_k} \text{Numerator}}$$

C. Drop insignificant terms to reduce the combined number of gaussians.

D. If

$$\left[ \phi_{k+1}(\mu_{k_j}) P_{k_j} \phi_{k+1}(\mu_{k_j})^T + Q_k \right] s^2 = s^2 P'_{k+1_i} < E, \quad (104)$$

form  $p_{x_{k+1}}(\xi_{k+1} | z_k, \dots, z_0)$  as

$$p_{x_{k+1}}(\xi_{k+1} | z_k, \dots, z_0) = \sum_{i=1}^{\eta'_{k+1}} \alpha'_{k+1_i} N(\underline{x}_{k+1} - \underline{a}_{k+1_i}, P'_{k+1_i}) \quad (105)$$

where:

$$\eta'_{k+1} = \eta_k, \quad \alpha'_{k+1_i} = \alpha_{k_i}$$

$$\underline{a}_{k+1_i} = \phi_{k+1}(\mu_{k_i})$$

$$P'_{k+1_i} = \phi_{k+1}(\mu_{k_i}) P_{k_i} \phi_{k+1}(\mu_{k_i})^T + Q_k$$

and go to step F. If (104) does not hold, go to step E.

E. Given a prespecified gaussian sum approximation to  $p(\underline{u}_k)$  of the form:

$$p_{u_k}(\underline{u}_k) = \sum_{n=1}^{q_k} \gamma_{k_n} N(\underline{u}_k - \underline{u}_{k_n}, Q_{k_n}) \quad (106)$$

which leads to

$$p_{o_{u_k}}(\underline{x}_{k+1} | \underline{x}_k) = \sum_{n=1}^{q_k} \gamma_{k_n} N \left[ \underline{x}_{k+1} - \phi_{k+1}(\underline{\mu}_{k_j}) - \phi_{k+1}(\underline{\mu}_{k_j})(\underline{x}_k - \underline{\mu}_{k_j}) - \underline{\omega}_{k_n}, Q_{k_n} \right] \quad (107)$$

where  $\phi_{k+1}$  is linearized about  $\underline{\mu}_{k_j}$ . With this approximation, if each  $k+1$  stage covariance is such that

$$s^2 P'_{k+1} = s^2 \left[ \phi_{k+1}(\underline{\mu}_{k_j}) P_{k_j} \phi_{k+1}(\underline{\mu}_{k_j})^T + Q_{k_n} \right] < E, \quad (108)$$

form  $p_{x_{k+1}}(\underline{x}_{k+1} | \underline{z}_k, \dots, \underline{z}_0)$

$$p_{x_{k+1}}(\underline{x}_{k+1} | \underline{z}_k, \dots, \underline{z}_0) = \sum_{i=1}^{\eta'_{k+1}} \alpha'_{k+1_i} N(\underline{x}_{k+1} - \underline{a}_{k+1_i}, P'_{k+1_i}) \quad (109)$$

where

$$\eta'_{k+1} = \eta_k q_k,$$

$$\alpha'_{k+1_i} = \alpha_{k_j} \gamma_{k_n},$$

$$\underline{a}_{k+1_i} = \phi_{k+1}(\underline{\mu}_{k_j}) + \underline{\omega}_{k_n},$$

$$P'_{k+1_i} = \phi_{k+1}(\underline{\mu}_{k_j}) P_{k_j} \phi_{k+1}(\underline{\mu}_{k_j})^T + Q_{k_n},$$

and where the index of summation comes from combining the terms of  $p(\underline{x}_k | \underline{z}_k)$  and  $p(\underline{x}_{k+1} | \underline{x}_k)$  such that  $i = j + (n-1) \eta_k$ . Next go to step F. If (108) is not valid go to step G.

F. Update the time index  $k$  to  $k+1$  and return to step B.

G. Form  $p_{x_{k+1}}(\xi_{k+1} | z_k, \dots, z_0)$  as in step C, (105) and then update the time index  $k$  to  $k+1$  and return to step A.

The basic procedure of Table 1 has been used in an example in Chapter VI, where certain simplifications have resulted from symmetry (see Chapter VI). For more details concerning other special cases and certain accuracy discussions the reader is encouraged to consult Alspach [1].

## E. Other Computational Methods

### 1. Fourier Series Expansions

Whenever the conditional densities have compact support and are periodic (for example, in the cyclic version of the problem in Chapter VII) it is frequently possible to expand both sides of the Bayes Law equation in a Fourier Series. Although such an expansion would always be theoretically possible, its advantage would not be particularly good unless it is possible to obtain (analytically) a closed form for the updating of the coefficients through Bayes Law. It is also important that proper consideration be given to the problem of truncation of the series, since (as generally will be the case) the coefficients for the new expansion will be given as an infinite sum of the previous coefficients. For example, we derive in Chapter VII (Appendix D) an expression of the form

$$a_{m\ell}^n = \frac{\tilde{m}_\ell \sum_{\alpha} S_{m-\alpha} a_{\alpha, \alpha-\ell}^{n-1}}{\sum_{\alpha} S_{-\alpha} a_{\alpha\alpha}^{n-1}}, \quad (110)$$

where  $a_{m\ell}^n$  are the (two-dimensional) Fourier coefficients for expanding the conditional probability density at time  $n$ , and the  $\tilde{m}_\ell$  and  $S_\alpha$  are coefficients resulting from the derivation of the recursion formula. The recursion involves two infinite sums. Of course if the density is uniform then only one coefficient will be non-zero. Thus, in contrast to the Gauss-Hermite expansion, the easiest situation to represent by a Fourier Series is the noise-saturated case.

For low-noise applications, however, when the density contains a maximum amount of structure (i.e., information) then the Fourier Series will begin to contain many high frequency components so that the recursion formula may lead to high frequency instabilities. We are currently studying methods for reducing the impact of the problem on the one hand, and for quantifying its effects on the other hand. We will report these results at a later date.

## 2. Spline Functions

It turns out that one more generalization of the least-squares ( $L_2$ ) approach of Center [6] leads to the method of interpolating splines, as revealed by Weinert and Kailath [12]. The additional generalization is the inclusion of constraints in the minimization process. The constraints take the form of knots where the values of the function are assumed known. Typically the knots are chosen so that the spline residuals are minimized, but the optimal choice of locations can often be a difficult proposition

In a sense the spline interpolation is a combination of the point-mass method with the least-squares fitting of polynomials. Thus it may be expected to include the advantages of both methods. Preliminary experiments by deFigueiredo and Jan [7], [10] appear to confirm this conjecture, although more definitive experiments are called for. DeFigueiredo and Jan have used multivariate fundamental splines ("B-splines") for the Bayes law calculation since "they are nonnegative; their integral over the reals is unity; and they give rise to a set of basis functions (called 'fundamental functions') with minimum support,

in terms of which a continuous nonnegative function is approximated by interpolation, as a nonnegative function" [8].

Using cubic polynomials as basis functions, deFigueiredo and Jan have considered a recursion which begins at stage  $k$  with a given mesh and the basis functions. Then they compute the projections of the given measurement and state-increment probability densities onto the space spanned by the basis functions. Then by substitution of the resulting inner products into the Bayes recursion formula an expression for the aposteriori density function in terms of an infinite sum results. The next step depends on the application. If the old mesh is adequate for the new density, then an expansion is made and the process is repeated. If the old mesh no longer suffices then a new mesh is chosen, either deterministically (analogous to the predictive gridding of Bucy and Senne [6]) or iteratively (i.e., using a surface searching technique). In any event the spline technique involves both the computational difficulties and the advantages of the point-mass and least-squares series representations. The method promises, however, to provide superior performance to both methods given the same number of computations. An extensive computational study will be required to determine if this conjecture is true. We regret not having been able to undertake such an experiment.

## References

- [ 1 ] D.L. Alspach, "A Bayesian Approximation Technique for Estimation and Control of Time Discrete Stochastic Systems," Ph.D. Dissertation, University of California, San Diego, 1970.
- [ 2 ] R.S. Bucy, "Bayes Theorem and Digital Realizations for Non-Linear Filters," J. Astro. Sci. 17 (1969), 80-94.
- [ 3 ] R.S. Bucy, "Building and Evaluating Non-Linear Filters," To appear, Proc. Symp. on Appl. Math.; Stochastic Diff. Eqns., Amer. Math. Soc., April 1972.
- [ 4 ] R.S. Bucy and P.D. Joseph, Filtering for Stochastic Processes with Applications to Guidance, Wiley Interscience, New York, 1968.
- [ 5 ] R.S. Bucy, R.A. Geesey, and K.D. Senne, "A Passive Receiver Design via Nonlinear Filtering Theory," Proc. Third Hawaii International Conf. On System Sciences, Vol. I, 1970, 477-480.
- [ 6 ] R.S. Bucy and K.D. Senne, "Digital Synthesis of Nonlinear Filters," Automatica, 7 (1971), 287-298.
- [ 7 ] J.L. Center, "Practical Nonlinear Filtering of Discrete Observations by Generalized Least Squares Approximation of the Conditional Probability Distribution," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1971, 88-99.
- [ 8 ] R.J.P. de Figueiredo and Y.G. Jan, "Spline Filters," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1971, 127-138.
- [ 9 ] C. Hecht, "Synthesis and Realization of Nonlinear Filters," Ph.D. Dissertation, University of California, 1972.
- [10] F.B. Hildebrand, Introduction to Numerical Analysis, McGraw Hill, New York, 1956.
- [11] Y.G. Jan, Ph.D. Dissertation, Rice University, 1971.
- [12] J.T. Lo, "Finite Dimensional Sensor Orbits and Nonlinear Filtering," Ph.D. Dissertation, University of Southern California, 1969.
- [13] H.L. Weinert and T. Kailath, "Recursive Spline Interpolation and Least-Squares Estimation," submitted to Amer. Math. Soc., 1971.
- [14] N. Wiener, The Fourier Integral and Certain of Its Applications, Cambridge University Press, Cambridge, 1933 (Also: Dover, New York, 1958).

#### IV. Monte Carlo Analysis Techniques

##### A. Introduction

It is unfortunate that in most nonlinear filtering problems of interest the analytic performance evaluation of the estimates is as untractable as the derivation of the optimal estimates themselves. Naturally, in order to compare alternative estimates an error evaluation procedure of high quality is essential. Thus it is necessary to turn to Monte Carlo simulation. This section is devoted to describing the necessary experiments and to clearing up some of the misconceptions and misuses of the Monte Carlo method.

The term "Monte Carlo Simulation" is subjected to a great number of alternative and sometimes misleading interpretations. Some authors claim that significant conclusions can be made on the basis of averaging only a few random numbers. Others claim only that Monte Carlo simulation has proved that one technique is superior to another without any explanation of the type of experiment or the confidence associated with the conclusions. Other investigators, aware of the cost of such simulations, regret to say that, due to the prohibit expense, they were limited to 2-10 Monte Carlos, but then they come to conclusions completely inconsistent with that limitation. The result is that the Monte Carlo simulation has become the most prevalent form of cheating with statistics in the engineering literature. Consequently, we intend to reverse that trend by describing precisely what our simulations entailed and analyzing the confidence associated with our conclusions.

**Preceding page blank**

In addition, in an appendix, we present our random number generator, which is realizable on any binary computer, so that any reader can convince himself of our conclusions by precisely duplicating them if he chooses.

### B. Statistical Analysis

By Monte Carlo simulation we mean that, in order to estimate the moments of a random process generated by subtracting an estimate  $\hat{x}_n$  of a variable  $x_n$  from the true value, a large number of statistically independent realizations of the difference are generated and the statistic

$$\hat{\mu}_m = \frac{1}{N} \sum_{i=1}^N (\hat{x}_n^i - x_n^i)^m \quad (1)$$

is formed. If  $\mu_m$  is the  $m$ th moment of the difference  $\hat{x}_n^i - x_n^i$  for all  $i$ , then  $\hat{\mu}_m$  will have mean  $\mu_m$  as expected but what confidence is associated with the estimate, or, equivalently, how large should  $N$  be? The statistic (1) is asymptotically normal with mean and variance given by [2]

$$E[\hat{\mu}_m] = \mu_m, \quad (2)$$

and

$$\text{var} [\hat{\mu}_m] = \frac{\mu_{2m} - \mu_m^2}{N}. \quad (3)$$

Since the statistic is Gaussian we may compute the probability confidence bands for the estimates for large  $N$ . As an example, consider the probability (0.9974) that a gaussian deviate lies within three standard deviations of its mean. Thus, for large  $N$ , we claim with probability 0.9974 that

$$|\hat{\mu}_m - \mu_m| \leq 3 \sqrt{\frac{\mu_{2m} - \mu_m^2}{N}}. \quad (4)$$

An equivalent statement to (4) is

$$\frac{\hat{\mu}_m}{1 + 3 \sqrt{\frac{P_m - 1}{N}}} \leq \mu_m \leq \frac{\hat{\mu}_m}{1 - 3 \sqrt{\frac{P_m - 1}{N}}}, \quad (5)$$

where  $P_m = \frac{\mu_{2m}}{\mu_m^2}$ , provided  $\mu_m \neq 0$ .

Now the bounds in (5) are asymptotic results but we may still use them effectively if we know we have taken  $N$  large enough. One technique is to compute the sample distribution of  $\hat{\mu}_m$  (by performing  $M$  Monte Carlos of length  $N$ ) and test its normality with (say) a Chi-squared or Kolmogorov-Smirnov test. More reasonably, an upper bound on the value of  $P_m$  may be computed for use in (5) based on the samples. But it is definitely true that any confidence band of fewer than three standard deviations would be at best subject to misinterpretation since the associated probability for plus or minus one standard deviation is only 0.6836 and only 0.9544 for two. It is unfortunate, however, that the convergence of the bound (4) is extremely slow with  $N$ . If we compute the total width  $W(P_m, N)$  of the confidence band for  $n_\sigma$  standard deviations, we get

$$\frac{W(P_m, N)}{\hat{\mu}_m n_\sigma} = 2 \sqrt{\frac{P_m - 1}{N}}. \quad (6)$$

Examples of the normalized confidence widths are given in Table 1 for various values of  $P_m$  and  $N$ . It can be seen that for the three-standard deviation confidence width to be less than  $0.2 \hat{\mu}_m$ , and  $N$  of 2000 is needed if  $P_m = 3$ ,  $N = 5000$  is needed for  $P_m = 5$ , and  $N = 10000$  is needed

for  $P_m = 10$ . Thus, in order to separate two Monte Carlo results with high confidence (probability 0.9974) if the numbers lie within 10% of each other, an extremely large number of runs is required. Also, the table shows that absolutely nothing can be said with high confidence if fewer than 20 Monte Carlos are performed (with  $P_m = 3$ ) or 50 Monte Carlos (with  $P_m = 6$ ).

Table 1. The normalized standard deviation  $2\left(\frac{P-1}{N}\right)^{1/2}$  as a function of P and N

N	P = 1	2	3	4	5	6	8	10
1	0	2.00	2.83	3.46	4.00	4.47	5.29	6.00
2	0	1.41	2.00	2.45	2.83	3.16	3.74	4.24
5	0	.0894	1.26	1.55	2.24	2.00	2.37	2.68
10	0	.632	.894	1.10	1.26	1.41	1.67	1.90
20	0	.447	.632	.775	.894	1.00	1.18	1.34
50	0	.283	.400	.490	.566	.632	.748	.849
$10^2$	0	.200	.283	.346	.400	.447	.529	.600
$2 \times 10^2$	0	.141	.200	.245	.283	.316	.374	.424
$5 \times 10^2$	0	.0894	.126	.155	.224	.200	.237	.268
$10^3$	0	.0632	.0894	.110	.126	.141	.167	.190
$2 \times 10^3$	0	.0447	.0632	.0775	.0894	.100	.118	.134
$5 \times 10^3$	0	.0283	.0400	.0490	.0566	.0632	.0748	.0849
$10^4$	0	.0200	.0283	.0346	.0400	.0447	.0529	.0600
$2 \times 10^4$	0	.0141	.0200	.0245	.0283	.0316	.0374	.0424
$5 \times 10^4$	0	.00894	.0126	.0155	.0224	.0200	.0237	.0268
$10^5$	0	.00632	.00894	.0110	.0126	.0141	.0167	.0190
$2 \times 10^5$	0	.00447	.00632	.00775	.00894	.0100	.0118	.0134
$5 \times 10^5$	0	.00283	.00400	.00490	.00566	.00632	.00748	.00849
$10^6$	0	.00200	.00283	.00346	.00400	.00447	.00529	.00600
$2 \times 10^6$	0	.00141	.00200	.00245	.00283	.00316	.00374	.00424
$5 \times 10^6$	0	.000894	.00126	.00155	.00224	.00200	.00237	.00268
$10^7$	0	.000632	.000894	.00100	.00126	.00100	.00118	.00134

In case the hypothesized actual moment  $\mu_m$  is zero (for example: the odd moments of an even-valued density function), then the two-sided bound of (5) must be replaced (by setting  $\mu_m = 0$  in (4)) with

$$|\hat{\mu}_m| \leq n_\sigma \sqrt{\frac{\mu_{2m}}{N}} \quad (7)$$

Equation (7) will be used to test the accuracy of the odd moments of a gaussian random number generator below.

In addition to testing the accuracy of the overall statistic  $\hat{\mu}_m(N)$ , we are also interested in the allowable behavior of the cumulative average, whereby  $\hat{\mu}_m(N+1)$  is obtained from  $\hat{\mu}_m(N)$  by

$$\hat{\mu}_m(N+1) = \frac{N}{N+1} \hat{\mu}_m(N) + \frac{(\hat{X}_n^{N+1} X_n^{N+1})^m}{N+1} \quad (8)$$

Clearly the sample paths of the cumulative average will steadily approach its mean  $\mu_m$  with fluctuations gradually diminishing asymptotically to zero as predicted by the bounds (5) and (7). Thus if the sample path has been stored we may detect statistically questionable samples by placing a pair of confidence bands of (say) two standard deviations around the entire trajectory of  $\hat{\mu}_m(n)$ ,  $n=1, \dots, N$  using the assumed value of  $\mu_m$  equal to the asymptotic sample moment  $\hat{\mu}_m(N)$ . Frequently the trajectory of  $\hat{\mu}_m(N)$  will reveal programming errors whereby an occasional error is extremely large (for example - by dividing an estimate by zero unexpectedly). It is important to determine not only the asymptotic moment  $\hat{\mu}_m(N)$ , then, but also to estimate whether any of the individual samples deviates unusually far from  $\hat{\mu}_m(N)$ . A typical plot of a questionable sample path is given in Fig. 1. The behavior of the figure suggests a way in which bad samples (i.e., not taken from

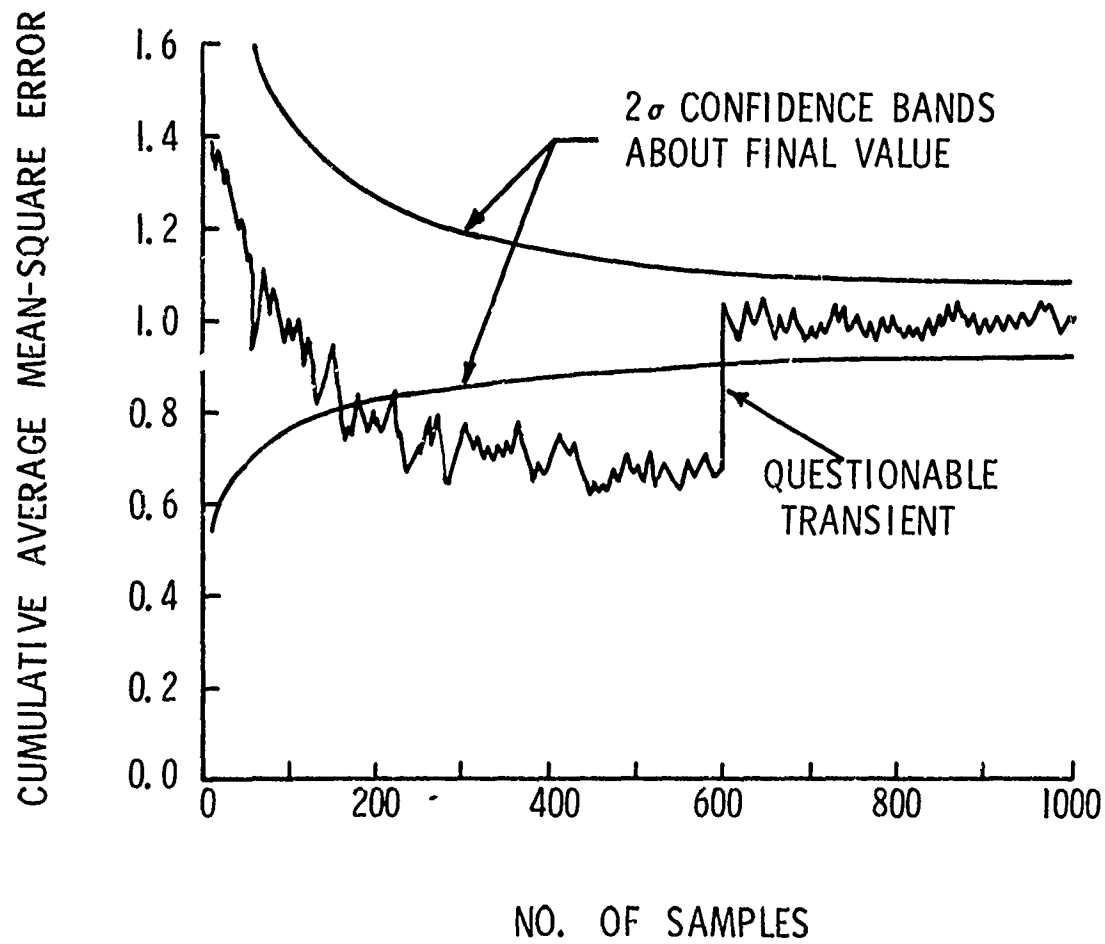


Fig. 1. Example of a Questionable Monte Carlo  
Cumulative Average Sample Path

the correct distribution) may be identified during the experiment. We simply compute the variance of the difference  $\hat{\mu}_m(N+1) - \hat{\mu}_m(N)$  as

$$\text{var}[\hat{\mu}_m(N+1) - \hat{\mu}_m(N)] \triangleq \text{var}[\delta\hat{\mu}_m(N)] = \frac{1}{(N+1)^2} \text{var}[\hat{\mu}_m(N)] \quad (9)$$

As an example of the use of (9) and Table 1, consider a situation where  $\hat{\mu}_m(100) = 3$  and we are interested in the three-standard-deviation allowable change at  $N=100$ . We observe that  $\text{var}[\hat{\mu}_m(N)] \approx \hat{\mu}_m^2(N) [W(P_m, N)/2]^2$ . Thus, from (9), selecting  $P_m = 3$ , we compute

$$\text{var}[\delta\hat{\mu}_m(100)] \approx \frac{\hat{\mu}_m^2(100) [W(3, 100)/2]^2}{101^2} = \frac{3^2 [(.283)/2]^2}{101^2} \approx (.00420)^2. \quad (10)$$

From (9) we see that the three-standard deviation cutoff for  $|\delta\hat{\mu}_m(100)|$  is .0126. If the magnitude  $\delta\hat{\mu}_m(N)$  ever exceeds the  $3\sigma$  value we have good reason to suspect an inconsistent data value. If at the point that such a deviation is encountered the offending number is flagged, set aside, and not included in the cumulative average. A post experiment analysis may be able to discover if indeed a programming error could be responsible.

### C. An Example

An excellent example for illustrating some of the above concepts is the testing of the gaussian random number generator which was used for all of our subsequent Monte-Carlo experiments. A description of the generator is given in complete detail in the appendix. The normal deviates were obtained in pairs by the following polar transformation of the uniform deviates:

let  $u_1, u_2$  be independent and distributed uniform on  $[0,1)$ , then

$$x_1 = (-\log_e u_1)^{1/2} \sin(2\pi u_2), \quad (11)$$

$$\text{and} \quad x_2 = (-\log_e u_1)^{1/2} \cos(2\pi u_2) \quad (12)$$

are independent and distributed gaussian with zero mean and unit variance.

First we make a quick check of the uniform generator alone, testing its mean and variance based on two independent Monte Carlos of  $10^5$  samples. Sequence One used the seed 735776465527<sub>8</sub> (see Appendix) and Sequence Two used the seed 311037552421<sub>8</sub>. The Monte Carlo results of central moments were as follows

$$\text{Sequence One} \quad \hat{\mu}_1(10^5) = .498931 - 0.5 \quad \hat{\mu}_2(10^5) = .0834604 \quad (13)$$

$$\text{Sequence Two} \quad \hat{\mu}_1(10^5) = .500517 - 0.5 \quad \hat{\mu}_2(10^5) = .0833716 \quad (14)$$

Now since  $\mu_1 = 0.0$ ,  $\mu_2 = 0.0833\dots$ ,  $\mu_3 = 0.03125$ , and  $\mu_4 = 0.0125$ , we may effectively use the bounds in (4) and (7) to bracket the allowable deviations. For this experiment we determine

$$|\hat{\mu}_1| = n_{\sigma_1} \sqrt{\frac{.0833}{10^5}} = n_{\sigma_1} (9.1287 \times 10^{-4}) \quad (15)$$

and

$$|\hat{\mu}_2 - \mu_2| \leq n_{\sigma_2} \sqrt{\frac{.0125 - (.0833)^2}{10^5}} = \frac{n_{\sigma_2}}{6} \sqrt{\frac{1}{5 \times 10^5}} = n_{\sigma_2} (2.3570 \times 10^{-4}) \quad (16)$$

The number of standard deviations of the two experiments may be computed by substituting the values from (13) into (15) and (14) into (16), giving

	$n_{\sigma_1}$	$n_{\sigma_2}$
Sequence One	1.171	0.537
Sequence Two	0.566	0.162

Note that all four results are consistent with the hypothesized distributions. Next we combine the results of both experiments by averaging, generating a sequence of twice the length, resulting in  $\hat{\mu}_1(2 \times 10^5) = .499724 - 0.5$  and  $\hat{\mu}_2(2 \times 10^5) = 0.083416$ . The calculations of (14) and (15) (with  $N = 2 \times 10^5$ ) now yield  $n_{\sigma_1} = 0.428$  and  $n_{\sigma_2} = 0.496$ , respectively, also consistent with the hypothesis that the numbers are taken from a distribution with mean 0.5 and variance 0.0833....

The above results were hypothesis testing results - i.e., given a hypothesized distribution, what is the likelihood that the given samples come from that distribution? We will return to the hypothesis problem shortly, but meanwhile, what if we had no hypothesis? We would then have to use a different approach to establish bounds on the true moments. We would take sample values to estimate  $\mu_2$ , for example, and use the bounds (5) to determine the allowable range for  $\mu_2$  based on the sample  $\hat{\mu}_2(2 \times 10^5)$ . Taking a nominal value of  $\mu_2$  from this range we could then use (7) to determine whether the assumption that  $\mu_1 = 0$  were reasonable. If  $\mu_1 = 0$  were not reasonable according to (7) we would finally return to (4) and determine a more plausible range for  $\mu_1$ . The result of the above procedure will be confidence intervals about each of the moments of the sampled distribution.

Turning now to the transformed deviates (hypothesized gaussian) we describe a more extensive set of experiments. Using the transformation (10) - (11) on the above two uniform sequences we calculated the first six moments of the samples based on Monte Carlos of length  $N = 5 \times 10^6$  for both starting conditions. The raw sampled moments are summarized in Table 2.

Table 2. Monte Carlo Moments of Gaussian Generator

	Sequence One	Sequence Two	Average	Theory
$\hat{\mu}_1$	0.000036	0.000052	0.000044	0
$\hat{\mu}_2$	0.999851	0.999315	0.999583	1
$\hat{\mu}_3$	-0.000808	-0.000639	-0.000724	0
$\hat{\mu}_4$	2.996913	2.995235	2.996074	3
$\hat{\mu}_5$	-0.011973	-0.013211	-0.012592	0
$\hat{\mu}_6$	14.967856	14.974867	14.9713615	15

First we determine if the averaged moments (column 3 in Table 2) are consistent with the hypothesized moments based on a Monte Carlo of length  $10^7$ . The results are given in Table 3, where the odd moments are evaluated with (7) and the even moments with (4). As can be seen from the table, all of the computed standard deviations are less than 1.3, leading to a high degree of consistency with the Gaussian hypothesis for the first six moments. In order to compute the standard deviations in Table 3, the moments  $\mu_8 = 105$ ,  $\mu_{10} = 945$ , and  $\mu_{12} = 10395$  were

Table 3. Testing the Hypothesis of Gaussian Moments

Odd Moments			Even Moments		
m	$\left(\frac{\mu_{2m}}{10^7}\right)^{1/2}$	$n_{\sigma_m}$	m	$\left(\frac{\mu_{2m} - \mu_m^2}{10^7}\right)^{1/2}$	$n_{\sigma_m}$
1	$3.1623 \times 10^{-4}$	0.1391	2	$4.4721 \times 10^{-4}$	0.9324
3	$1.2247 \times 10^{-3}$	0.5911	4	$3.0984 \times 10^{-3}$	1.2671
5	$9.7211 \times 10^{-3}$	1.2953	6	$3.1890 \times 10^{-2}$	0.8980

required. The standard deviations in the table reflect knowledge of the exact moments to almost four digits for the first three moments and to almost three digits for the fourth through sixth moments.

Of course any finite number of sampled moments may match the hypothesized moments and still the density may not be in fact gaussian. Thus a more extensive test is required to check the density hypothesis itself. Statistical tests such as the Chi-squared test [3] and the Kolmogorov-Smirnov test [4] have been devised to study this very problem. The Kolmogorov test is the most valuable for the problem at hand since it is both asymptotically efficient and distribution free, whereas the Chi-squared test is not. Basically the Kolmogorov test begins by constructing an empirical cumulative distribution function from the samples  $x_1$ , assumed ordered such that  $x_1 < x_2 < \dots < x_N$ . Assume the samples are taken from an unknown distribution  $U(x)$ , i.e.,  $\Pr(x_1 < x) = U(x)$ . Then construct the empirical distribution as follows:

$$\text{let } H(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (\text{heaviside step function}) \quad (17)$$

Then

$$F_N(x) \triangleq \frac{1}{N} \sum_{i=1}^N H(x-x_i), \quad (18)$$

where  $x_i$  are the ordered samples. Suppose we are interested in testing the hypothesis  $H_0$ , where

$$H_0 : U(x) = F(x) \text{ (given) } . \quad (19)$$

Kolmogorov turned the  $H_0$  problem into a threshold test of the following form. Determine the value of

$$K_N = (N)^{1/2} \sup_x |F(x) - F_N(x)|, \quad (20)$$

then reject  $H_0$  if  $K_N$  is too large. In order to determine how large  $K_N$  could be expected to be it was necessary to calculate the distribution of  $K_N$ . Kolmogorov originally proved that

$$\lim_{N \rightarrow \infty} \Pr(K_N < \lambda) = \Phi(\lambda) = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j+1} e^{-2j^2 \lambda^2} \quad (21)$$

Later, Massey [5] derived a recursive method for obtaining  $\Pr(K_N < \lambda)$  for all  $N$ . Table 4, taken from Massey's paper, illustrates how quickly the distribution converges to its asymptotic limit. If  $N$  is large (say 5000) we may use (21) as a quick test for  $K_N$ . For a plot of (21) and its derivative, see Fig. 2.

Table 4.  $\Pr(K_N < \lambda)$  from Massey [5]

$N \quad \lambda =$	0.9	1.0	1.1	1.2	1.3	1.4
10	.66	.78	.85	.91	.95	.97
20	.65	.77	.85	.91	.94	.97
30	.65	.76	.85	.90	.94	.96
40	.64	.76	.84	.90	.94	.96
50	.64	.75	.84			
60	.63	.75	.84			
70	.63	.75	.83			
80	.63	.74				
$\infty$	.607	.730	.822	.888	.932	.960

In order to test the gaussian generator, then,  $M = 1000$  sample paths of length  $N = 5000$  were generated and the numbers  $K_N(i)$ ,  $i=1, \dots, M$

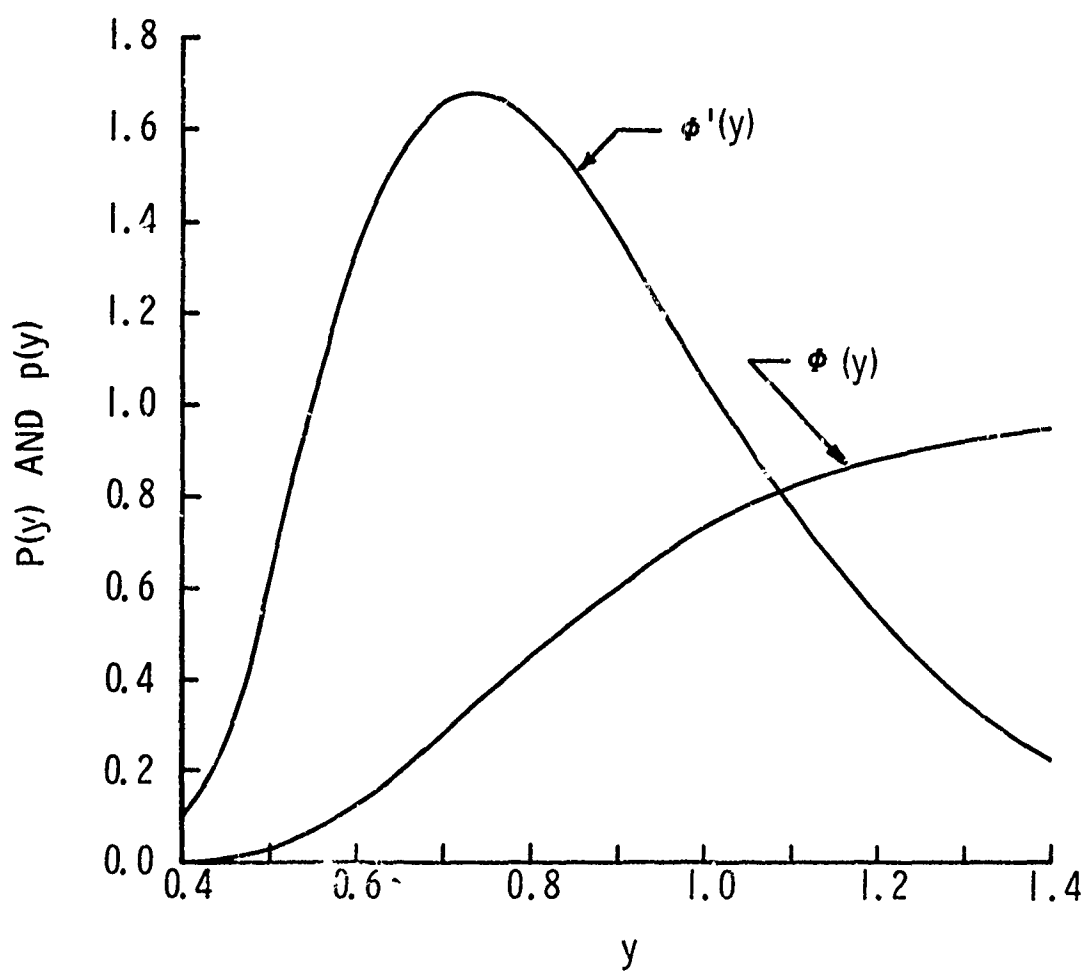


Fig. 2. Probability Density and Distribution  
of the Asymptotic ( $N \rightarrow \infty$ ) Kolmogorov Statistic

were calculated (where for convenience we assume  $K_N(1) < K_N(2) < \dots < K_N(M)$ ). Next an empirical distribution  $F_{K_N(M)}(\lambda)$  of the form (18) was determined and compared with (21). Table 5 shows some of the points compared for each sample path of the generator. Table 5 shows conclusively that not only is the maximum  $K_N(M)$  consistent with the hypothesis  $H_0$  (with  $F(x)$  gaussian) but the distribution of  $K_N(i)$  is consistent with  $H_0$ . Thus we have illustrated in effect a successful second order Kolmogorov test.

Table 5. Results of Kolmogorov Test

i/M	Sequence One		Sequence Two	
	$K_N(i)$	$\Phi[K_N(i)]$	$K_N(i)$	$\Phi[K_N(i)]$
0.1	0.5688	0.0973	0.5823	0.1132
0.2	0.6414	0.1949	0.6502	0.2084
0.3	0.6975	0.2845	0.7090	0.3038
0.4	0.7540	0.3795	0.7769	0.4179
0.5	0.8215	0.4904	0.8319	0.5068
0.6	0.8669	0.5599	0.8996	0.6067
0.7	0.9412	0.6616	0.9788	0.7065
0.8	1.0461	0.7761	1.0571	0.7862
0.9	1.1914	0.8831	1.2007	0.8881
1.0	1.8518	0.9979	1.8975	0.9985

We now know enough about the generator to state with high confidence that the distribution of the numbers is in fact Gaussian. We also will need to know if the successive numbers from the generator are independent.

Of course the test for independence is nothing more than to determine if the multivariate densities are all products of univariate Gaussians. Such a test would be prohibitively costly, due to the large number of samples required. Thus we will have to be satisfied by demonstrating that successive samples are uncorrelated. Accordingly, we compute a sampled correlation function defined by

$$\hat{\mu}_2(k) = \frac{1}{N-k} \sum_{i=k+1}^N (\hat{x}_n^i - \bar{x}_n^i)(\hat{x}_n^{i-k} - \bar{x}_n^{i-k}) \quad (22)$$

$k = 0, 1, \dots$ . The sampled correlation function just reduces to the autocorrelation  $\hat{\mu}_2$  of (1) for  $k = 0$ . Using the two sample functions of the generator we have obtained values of (21) for  $k = 1, \dots, 4$ . Refer to Table 6 for these experimental results.

Table 6. Sampled Correlation Function

	Sequence One	Sequence Two	Average
$k$	$\hat{\mu}_2(k)$	$\hat{\mu}_2(k)$	$\hat{\mu}_2(k)$
1	-0.000588	-0.000131	-.000360
2	-0.000031	0.000160	.000065
3	-0.000790	0.000568	-.000111
4	0.000588	-0.000129	.000230

Now the variance of (21) is given by

$$\text{var} [\hat{\mu}_2(k)] = \frac{\mu_2^2}{N-k} \quad (23)$$

Thus we may determine the number of standard deviations associated with the (say) the average of the correlations in Table 7, whereby  $N = 10^7$ .

The results are given in Table 7, where we see that uncorrelatedness is a highly consistent hypothesis for the generator, since the number of standard deviations is consistently less than 1.2.

Table 7. Uncorrelatedness Test

k	$\sigma = \left( \frac{\mu_2}{N-k} \right)^{1/2}$	$n_\sigma = \frac{ \hat{\mu}_2(k) }{\sigma}$
1	$3.1623 \times 10^{-4}$	1.1384
2	$3.1623 \times 10^{-4}$	0.2055
3	$3.1623 \times 10^{-4}$	0.3510
4	$3.1623 \times 10^{-4}$	0.7273

#### D. Conclusions

It has been seen that a systematic use of standard statistical results can lead to a qualitative level of confidence associated with a Monte Carlo analysis. Since the latter is of such great importance in evaluating the performance of nonlinear estimators, it goes almost without saying that no Monte Carlo results should be published without a high confidence assessment of their accuracy.

Naturally an important characteristic of simulations is reproducibility. Thus the machine-independent random number generator analyzed above will be of great service to those engaged in comparative analyses of several candidate filters, since the simulations need not necessarily be run on the same computer.

The two starting numbers provided for the random number generator in this paper yield two independent sequences, both of which are useful

in stochastic systems simulation. Furthermore, the characteristics of the initial segments of the two sequences are sufficiently different so as to provide complementary tests of an estimator with only two relatively short Monte Carlos.

## References

- [1] M. Abramowitz, and I. A. Stegun, Handbook of Mathematical Functions, Dover, New York, 1965.
- [2] R. S. Bucy, "Building and Evaluating Non-Linear Filters," Proceedings of the Symposium on Applied Mathematics; Stochastic Differential Equations, American Mathematics Society, April 1972.
- [3] H. Cramer, Mathematical Methods of Statistics, Princeton University Press, Princeton, 1951, pp 416-451.
- [4] J. L. Doob, "Heuristic Approach to the Kolmogorov-Smirnov Theorems," Annals of Mathematical Statistics, 20 (1949), 393-403.
- [5] F. J. Massey, Jr., "A Note on the Estimation of a Distribution Function by Confidence Limits," Annals of Mathematical Statistics, 21 (1950), 116-119.

## Appendix. A Machine Independent Random Number Generator

### A. Introduction

A continually recurring problem associated with Monte Carlo simulations on digital computers is the lack of reproducibility of results. The simulation performed by one researcher on one computer must essentially be believed by others without access to the same machine. This problem arises primarily from the fact that (1) generally the random number generator used on a given system has been optimized for speed and efficiency by system programmers, (2) the generator depends in subtle ways on the particular idiosyncrasies of the machine, and (3) rarely, if ever, does sufficient documentation exist so that the user can unravel the secrets of the method employed. This is not to say that such generators are not reliable but only that they are not reproducible. Thus it is frequently impossible for various researchers to compare their results for the same problem on different machines. It is the purpose of this appendix to propose a solution to the problem - to sacrifice speed and efficiency for reproducibility, independent of the word-length of the binary machine employed. The generation method is simply a congruence method [1] with the factors appropriately segmented so that numerical overflows can be computed without hardware overflows, since the latter can lead to machine-dependent results. It turns out that a convenient overall length for the equivalent random numbers is 36-bits, since such numbers can be evenly partitioned into 1, 2, 3, 4, 6, or 9 pieces and yet the cycle-length of the generator is

adequate for almost any conceivable experiment. It is demonstrated in this article that simple FORTRAN-II-compatible coding exists for the multipart generators so that reproducibility can be effectively achieved with no special-purpose coding, regardless of the machine. It is important to keep in mind here that speed is not the design criterion which is being considered in this article. However, many shortcuts could be made by using machine coding for the generators. An example of the 36-bit generator is given with test results for two starting numbers. The reader is invited to choose the generator having the appropriate number of segments for his computer and reproduce exactly the results contained here - to demonstrate the purpose of the paper. It is clear that the same segmentation principle will work on any deterministic generator which is defined by standard mathematical transformations.

### I. The Congruence Method

It is desired to produce a long sequence of random variates  $\{x_n\}$  which can pass any standard test for randomness and are distributed essentially uniformly on the interval  $[0,1)$ . A common method involves the calculation, for some length  $m$ , a sequence of integers  $\{l_n\}$  between 0 and  $2^m-1$  by modular arithmetic and obtaining the desired real variables  $x_n$  by dividing the  $l_n$  by  $2^m$ . For example, let  $l_{n+1}$  be obtained from  $l_n$  by the operation

$$l_{n+1} = (a l_n + b) \bmod 2^m \quad (A-1)$$

While many combinations for  $a$ ,  $b$ , and  $m$  have been proven successful [1], one of the more common combinations involves letting  $b = 0$  and taking  $a$  to be of the form  $r2^s+1$ , where  $s \geq 2$ . The multiplier  $r$

need only be chosen so that the number of significant bits in  $r2^s \pm 1$  is approximately  $m$ . In this way each application of (A-1) will force  $l_{n+1}$  into the next cycle, leaving little or no correlation between adjacent numbers. For this article we shall choose  $a = 8r - 1$ , or any  $m$ -bit number with the last three bits equal one. Thus we are left with the choice of  $m$ .\*

The common choice for the word-length of the  $l_n$  is the machine word length. Such a choice automatically makes the generator machine dependent, since the modular arithmetic contained in (A-1) will usually result in arithmetic overflows, the result of which is not predictable in general for all machines. Consequently we will consider segmenting the numbers into the form ( $m$  is evenly dividable by  $q$ )

$$l_n = l_n^1 2^{\frac{m(q-1)}{q}} + l_n^2 2^{\frac{m(q-2)}{q}} + \dots + l_n^q 2^0 \quad (A-2)$$

If we then segment  $a$  into the analogous pieces

$$a = a^1 2^{\frac{m(q-1)}{q}} + \dots + a^q 2^0, \quad (A-3)$$

we may perform the operation (A-1) as follows: First we compute

$$\begin{aligned} a l_n &= a^1 l_n^1 2^{\frac{2m(q-1)}{q}} \\ &+ (a^1 l_n^2 + a^2 l_n^1) 2^{\frac{2m(q-2)}{q}} + \dots \\ &+ (a^{q-1} l_n^q + a^q l_n^{q-1}) 2^{\frac{2m}{q}} + a^q l_n^q 2^0 \end{aligned} \quad (A-4)$$

---

\* See [1] for an equivalent discussion for decimal machines. The proposed technique applies only to binary machines.

Next we observe that all but the last half of the center term and the rest of the lower order terms are eliminated when we take the modulus relative to  $2^m$ , leaving

$$\begin{aligned}
 l_{n+1} = & \left\{ (a^1 l_n^q + a^2 l_n^{q-1} + \dots + a^q l_n^1) \bmod 2^{\frac{m}{q}} \right\} 2^{\frac{2m}{q} \lfloor \frac{q}{2} \rfloor} \\
 & + (a^2 l_n^q + a^3 l_n^{q-1} + \dots + a^q l_n^2) 2^{\frac{2m}{q} (\lfloor \frac{q}{2} \rfloor - 2)} \\
 & + \dots + a^q l_n^q 2^0, \tag{A-5}
 \end{aligned}$$

where  $[\cdot]$  is the usual bracket (integer part) function. Finally, we see that the last half of the last term in (A-5) is  $l_{n+1}^q$ , the remainder of the last term plus the last half of the next to last term is  $l_{n+1}^{q-1}$ , etc. We summarize the algorithm involved to identify the pieces of  $l_{n+1}$  in Fig. A-1, where it is assumed that  $a^k$  and  $l_n^k$  are defined from the previous operation or initially by a suitable seed.

We see from studying (A-5) that the maximum precision required to accomplish the update is only  $2q$ -bits plus the necessary carry-over bits to add up  $q$  numbers of length  $\frac{2m}{q}$  and one number of length  $\frac{m}{q}$ . This exact number of bits is  $p = \lceil \log_2 \{q(2^{\frac{m}{q}} - 1)^2 + 2^{\frac{m}{q}} - 1\} \rceil + 1$ , where the bracket function has again been used again to denote the largest integer.

### B. An Example

In this section we introduce a specific example of the generator for  $m = 36$ , which is chosen since it is sufficiently long, and yet evenly dividable by  $q = 2, 3, 4, 6, 9$ , etc. Thereby providing

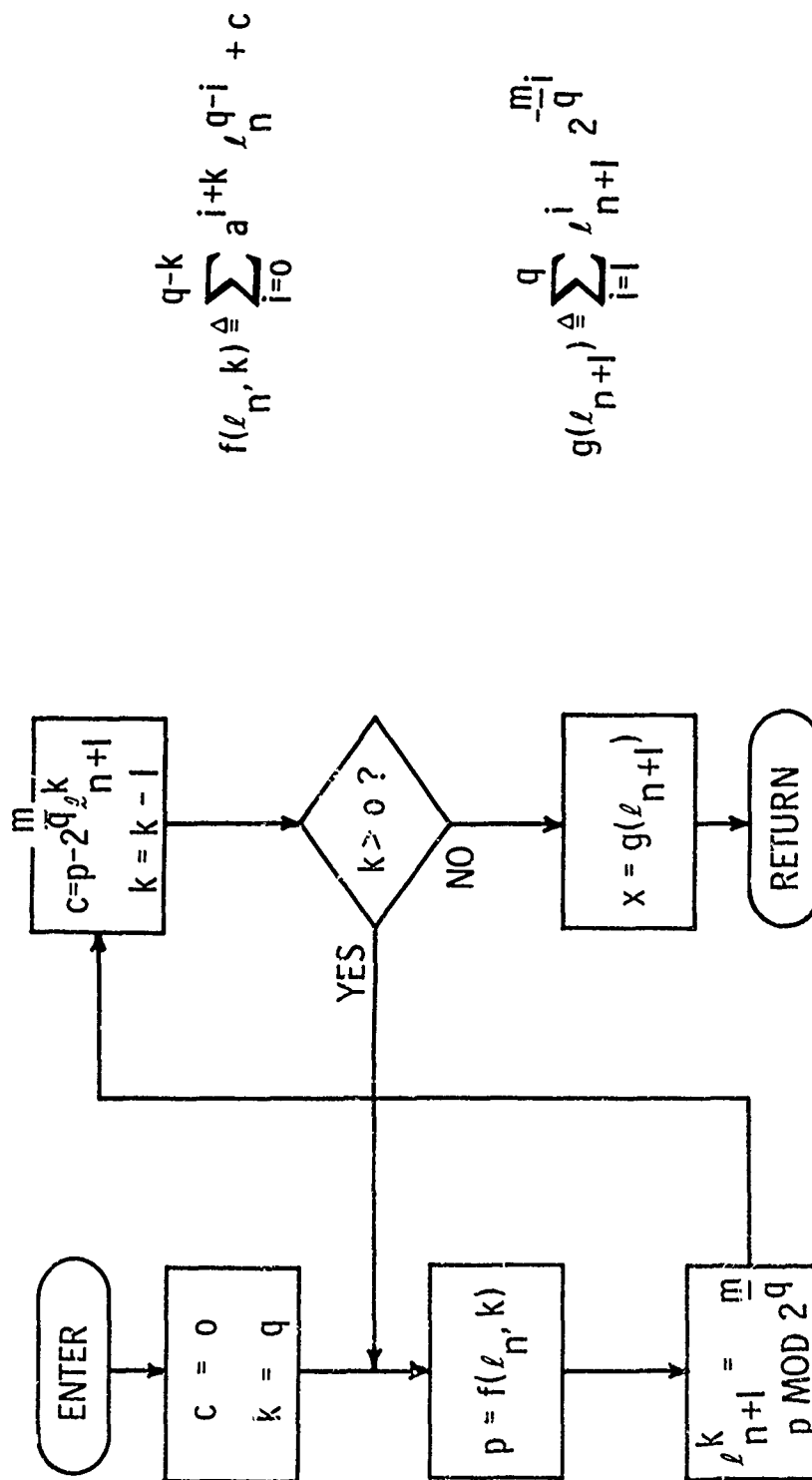


Fig. A-1. Algorithm for Partitioned Uniform Generator

significant flexibility. The choice of  $q^*$  is determined primarily so that the operation (A-5) does not lead to integer overflows on the machine concerned. Table A-1 summarized the hardware requirements for several combinations.

Table A-1. Partition Requirements  
for  $m = 36$  bit random numbers

$q = \text{No. of Pieces}$	$\frac{m}{q} = \text{Bits per Piece}$	Bits/(positive) integer
1	36	72
2	18	37
3	12	26
4	9	20
6	6	15
9	4	11
12	3	10
18	2	8
36	1	6

The number of carry bits increases at such a rate that it is not very efficient to divide the generator into more than about six or nine parts, but the principle remains the same regardless.

In order to initialize the generator it is necessary to provide  $q$  pieces of a suitable seed (i.e., one that leads to a reliable sequence). Table A-2 lists the multiplier  $a = 735776465527_8$  and the initial

---

\* Of course the individual pieces need not necessarily be the same length but this choice makes the algorithm simplest to code.

numbers  $1_0 = a$  (Sequence One and  $1_0 = 311037552421_8$  (Sequence Two) for several different dichotomizations.

Table A-2. Partition Examples

No. of Pieces	$a(1_0$ for Sequence One)	$1_0$ for Sequence Two
2	$735776_8 = 244734_{10}$	$311037_8 = 102943_{10}$
	$465527_8 = 158551_{10}$	$552421_8 = 185617_{10}$
3	$7357_8 = 3823_{10}$	$3110_8 = 1608_{10}$
	$7646_8 = 4006_{10}$	$3755_8 = 2029_{10}$
	$5527_8 = 2903_{10}$	$2421_8 = 1297_{10}$
4	$735_8 = 477_{10}$	$311_8 = 201_{10}$
	$776_8 = 510_{10}$	$037_8 = 031_{10}$
	$465_8 = 309_{10}$	$552_8 = 362_{10}$
	$527_8 = 343_{10}$	$421_8 = 273_{10}$
6	$73_8 = 59_{10}$	$31_8 = 25_{10}$
	$57_8 = 47_{10}$	$10_8 = 08_{10}$
	$76_8 = 62_{10}$	$37_8 = 31_{10}$
	$46_8 = 38_{10}$	$55_8 = 45_{10}$
	$55_8 = 45_{10}$	$24_8 = 20_{10}$
	$27_8 = 23_{10}$	$21_8 = 17_{10}$

Table A-3 contains the first 250 octal numbers resulting from Sequence One (equivalent to the seed  $1_0 = 1$  with one number discarded)

and Table A-4 contains the initial 250 octal numbers from Sequence Two. The statistics for these two sequences is given in the main part of this article. The cycle length of the generator may be calculated theoretically as follows: The next to last octal digit repeats every 8 steps, the third from last every 64 steps, etc., so that the first digit repeats every  $8^{11} = 2^{33} \approx 8.59 \times 10^9$ . In practice, however, the maximum cycle can only be obtained for certain seeds, thus a check must be made to guarantee that the cycle is sufficiently long. Such a test has been run on the two given starting numbers, confirming a cycle length of greater than  $10^7$  for both. The repeat characteristics for the initial segment of each sequence is given in Table A-5.

Table A-3. Initial Sample Path-- Sequence One

0/	735776465527	310473353621	167041756507	473777560041	712641373067
	546011324661	742044723047	640355342101	062221556427	526574737721
10/	605007105407	133551226141	350232117767	743450114761	305721405747
	124027314201	037341537327	474341134021	313250724307	350037704241
20/	176553134667	434502115061	430116160647	263655676301	757225410227
	767051140121	101236233207	560164372341	042233441567	652470325161
30/	003422023547	655044670401	716424151127	276125754221	563076232107
	717370070441	646162236467	445053745261	157523756447	353052272501
40/	004604602027	613250400321	515637721007	265173076541	124366323367
	051575575361	671013001347	575260304601	454316122727	042241634421
50/	715232077707	455020314641	045756700267	144316635461	150756114247
	120347726701	152627333627	431002700521	666303746607	046407042741
60/	372342545167	227000105561	355564317147	271101761001	350107434527
	337314554621	116564305507	775360601041	720430702067	241462565661
70/	122324612047	374157423101	707405425427	307500240721	424702334407
	001636347141	775630326767	612307455761	367006174747	326541475201
80/	336670306327	015736535021	570205053307	777441125241	146450243667
	605537556061	510277647647	005511157301	157007057227	340550641121
90/	221123262207	706711713341	266117450567	775734066161	170170412547
	316227251401	633530520127	350337555221	330604161107	264051511441
100/	226525145467	633535606261	465147245447	205774753501	543324051027
	317604301321	741056550007	467421317541	351100132367	115505536361
110/	142603170347	032153065601	327540271727	113327635421	015051626707
	135622135641	375327407267	520464676461	721203203247	230422607701
120/	524644402627	042213001521	110414375607	527374763741	712242154167
	467414246561	365736306147	666544742001	272007403527	265316755621
130/	700055634507	401342622041	445347211067	763355026661	642713501047
	132422504101	641260274427	607664541721	631244563407	554424470141
140/	373255535767	333470016761	277101763747	442014656201	627606055327
	106715136021	252510202307	545643346241	207274352667	265216217061
150/	376370336647	353204440301	525057526227	721731342121	463257311207
	776340234341	613632457567	673720627161	631546001547	534352632401
160/	364024067127	502332356221	223461110107	023534132441	611617054467
	651040447261	406301534447	553760434501	052132320027	364021202321
170/	404344377007	726750040541	371240741367	674336477361	657002357347
	750206646601	177751440727	614376636421	517640355707	125624756641
180/	201227116267	303653737461	252737272247	777736470701	610550451627
	403544102521	210374024607	401663704741	057370563167	443151407561
190/	166317275147	711550723001	432476352527	653502156621	702116163507
	470725643041	260414520067	334470267661	652411370047	314326565101
200/	606622143427	470332042721	711256012407	714713611141	207731744767
	630171357761	665204552747	207031037201	721112624327	710254537021
210/	333362331307	213646567241	510047461667	254315660061	621370025647
	370540721301	750417175227	353573043121	136662340207	707667555341
220/	312174466567	445426370161	175732370547	450437213401	336506436127
	255106157221	232505037107	737217553441	146437763467	716184310261
230/	272143023447	050005115501	440027567027	231117103321	156701226007
	304377561541	034030550367	467110440361	064610546347	470403427601
240/	254151607727	306426637421	414376104707	606030577641	030455625267
	677064000461	515002361247	427513351701	515343520527	336136203521

Table A-4. Initial Sample Path - Sequence Two

0/	311037552421	222760501707	553141172641	645504742267	421051453461
	473364616247	332563504701	566217475627	113326416521	436576550607
10/	542451520741	731701007167	410550523561	125147221147	002327337001
	547717776527	273546072621	006643307507	135125057041	311217344067
20/	675431003661	234703714047	341176601101	446056167427	732017356721
	053565536407	770664425141	113267170767	143633473761	024401476747
30/	502132453201	610621250327	320523453021	336714455307	717531003241
	106377305667	335621374061	512727351647	160233735301	522640221227
40/	473763357121	632477064207	377623371341	747756712567	703733504161
	023674314547	676663627401	551702062127	062360073221	672044163107
50/	663364767441	333714607467	154633024261	077747347447	421123131501
	416435613027	514612417321	061222752007	251116375541	205237774367
60/	365060554361	124517472347	542553043601	024232233727	655503553421
	505142230707	273261013641	466057451267	107475514461	104253705247
70/	517054365701	431336544627	411734517521	103451177607	150555441741
	547002416167	045242664561	046163210147	632010320001	677121745527
80/	260526273621	473076636507	534025100041	357537653067	455201244661
	356334603047	667257662101	341433036427	222761657721	263271765407
90/	764170546141	076516377767	344472034761	445737265747	123023634201
	317441017327	114060054021	544561604307	314251224241	621225414667
100/	676556035061	642462040647	731145216301	603012670227	051322060121
	760375113207	103367712341	053673721567	444776245161	077113703547
110/	376025210401	330477431127	062330674221	230663112107	004765410441
	621410516467	577013665261	223143636447	505724612501	507746062027
120/	460605320321	370652601007	220163116541	274202603367	072367515361
	253160661347	615224624601	633345402727	670130554421	674272757707
130/	300601634641	524761160267	723142555461	700451774247	220606246701
	050344613627	456423620521	130172626607	526162362741	353333025167
140/	733056025561	361406177147	565032301001	007112714527	323667474621
	302101165507	174326121041	056207162067	402172505661	771074472047
150/	117001743101	712476705427	312205160721	450445214407	523175667141
	042774606767	436651375761	747304054747	253456015201	470647566327
160/	435775455221	652775733307	566572445241	536002523667	465133476061
	731123527647	027717477301	602454337227	315341561121	021542142207
170/	207035233341	472437730567	577562006161	660142272547	251127571401
	573464000127	321062475221	605651041107	660367031441	422633425467
180/	567015526261	313047125447	604567273501	722345331027	443461221321
	111351430007	463330637541	567574412367	731617456361	275231050347
190/	550037405601	510447551727	711536555421	557372506707	754323455641
	551211667267	264630616461	007157063247	460601127701	352241662627
200/	732173721521	223563255607	534070303741	616112434167	553012166561
	316040166147	144415262001	106672663527	206211675621	622652514507
210/	257230142041	554005471067	275404746661	623743361047	671165024101
	650231554427	040711461721	102267443407	506704010141	237302015767
220/	323351736761	757657643747	634651176201	113445335327	447274056021
	050561062307	477714666241	023506632667	502132137061	505474216647
230/	275332760301	230405006227	511042262121	065156171207	573403554341
	473032737567	485066547161	575777661547	221173152401	131637347127
240/	157375276221	17005770107	246771452441	506605334467	325640367261
	676461414447	736472754501	365033600027	706216122321	732117257007

Table A-5. Repeat Characteristics  
of the Generator for each Sequence

Cycle	Sequence One	Sequence Two
0	<u>735776465527</u> <sub>8</sub>	<u>311037552421</u> <sub>8</sub>
8 <sup>1</sup>	<u>062221556427</u> <sub>8</sub>	<u>113326416521</u> <sub>8</sub>
8 <sup>2</sup>	<u>650107434527</u> <sub>8</sub>	<u>655503553421</u> <sub>8</sub>
8 <sup>3</sup>	<u>617512155527</u> <sub>8</sub>	<u>514633562421</u> <sub>8</sub>
8 <sup>4</sup>	<u>65527</u> <sub>8</sub>	<u>52421</u> <sub>8</sub>
8 <sup>5</sup>	<u>465527</u> <sub>8</sub>	<u>552421</u> <sub>8</sub>
8 <sup>6</sup>	<u>6465527</u> <sub>8</sub>	<u>7552421</u> <sub>8</sub>

The final concern in describing the generator is to give examples of coding the generator. These examples, written in FORTRAN II-compatible code, assume no special hardware characteristics except that only the applicable number of pieces has been selected. It is further assumed that the subroutine remains core-resident so that all locally defined variables remain unchanged between calls. If the subroutine must be dynamically reloaded on call (either because of load on call restrictions or virtual core) then the locally defined variables must be stored globally in COMMON. We leave this modification to the user. Table A-6 contains coding examples for two, three, four, and six-piece generators. No optimization has been done.

Table A-6. FORTRAN-II Coding Examples

## Two-Piece Generator

```
FUNCTION RANF(NS,MODE)
  DIMENSION NS(1),NC(2)
  IF (MODE) 10,100,10
10 M1=244734
   M2=158551
   N1=NS(1)
   N2=NS(2)
   T1=2.**(-18)
   T2=2.**(-36)
   MP=2**18
100 DO 200 I=1,2
    GO TO (110,120),I
110 K=M2*N2
    GO TO 190
120 K=M1*N2+M2*N1+KD
190 KL=K/MP
200 NC(1)=K-KL*MP
    N1=NC(2)
    N2=NC(1)
    XN1=N1
    XN2=N2
    RANF=XN1*T1+XN2*T2
    RETURN
END
```

## Three-Piece Generator

```
FUNCTION RANF(NS,MODE)
DIMENSION NS(1),NC(3)
IF (MODE) 10,100,10
10 M1=3823
   M2=4006
   M3=2903
   N1=NS(1)
   N2=NS(2)
   N3=NS(3)
   T1=2.**(-12)
   T2=2.**(-24)
   T3=2.**(-36)
   MP=2**12
100 DO 200 I=1,3
    GO TO (110,120,130),I
110 K=N3*M3
    GO TO 190
120 K=N3*M2+N2*M3+KD
    GO TO 190
130 K=N3*M1+N2*M2+N1*M3+KD
190 KD=K/MP
200 NC(I)=K-KD*MP
    N1=NC(3)
    N2=NC(2)
    N3=NC(1)
    XN1=N1
    XN2=N2
    XN3=N3
    RANF=XN1*T1+XN2*T2+XN3*T3
    RETU
END
```

## Four-Piece Generator

```

FUNCTION RANF(NS,MODE)
DIMENSION NS(1),NC(4)
IF (MODE) 10,100,10
10 M1=477
   M2=510
   M3=309
   M4=343
   N1=NS(1)
   N2=NS(2)
   N3=NS(3)
   N4=NS(4)
   T1=2.**(-9)
   T2=2.**(-18)
   T3=3.**(-27)
   T4=4.**(-36)
   MP=2**(9)
100 DO 200 I=1,4
    GO TO (110,120,130,140),I
110 K=N4*M4
    GO TO 190
120 K=N4*M1+N3*M4+KD
    GO TO 190
130 K=N4*M2+N3*M3+N2*M4+KD
    GO TO 190
140 K=N4*M1+N3*M2+N2*M3+N1*M4+KD
190 KD=K/MP
200 NC(I)=K-KD*MP
    N1=NC(4)
    N2=NC(3)
    N3=NC(2)
    N4=NC(1)
    XN1=N1
    XN2=N2
    XN3=N3
    XN4=N4
    RANF=XN1*T1+XN2*T2+XN3*T3+XN4*T4
    RETURN
END

```

## Six-Piece Generator

```

FUNCTION RANF(NS,MODE)
DIMENSION NS(1),NC(6)
IF (MODE) 10,100,10
10 M1=59
   M2=47
   M3=62
   M4=38
   M5=45
   M6=23
   N1=NS(1)
   N2=NS(2)
   N3=NS(3)
   N4=NS(4)
   N5=NS(5)
   N6=NS(6)
   T1=2.**(-6)
   T2=2.**(-12)
   T3=2.**(-18)
   T4=2.**(-24)
   T5=2.**(-30)
   T6=2.**(-36)
   MP=2**(6)
100 DO 200 I=1,6
    GO TO (110,120,130,140,150,160),I
110 K=N6*M6
    GO TO 190
120 K=N6*M5+N5*M6+KD
    GO TO 190
130 K=N6*M4+N5*M5+N4*M6+KD
    GO TO 190
140 K=N6*M3+N5*M4+N4*M5+N3*M6+KD
    GO TO 190
150 K=N6*M2+N5*M3+N4*M4+N3*M5+N2*M6+KD
    GO TO 190
160 K=N6*M1+N5*M2+N4*M3+N3*M4+N2*M5+N1*M6+KD
190 KD=K/MP
200 NC(I)=K-KD*MP
    N1=NC(6)
    N2=NC(5)
    N3=NC(4)
    N4=NC(3)
    N5=NC(2)
    N6=NC(1)
    XN1=N1
    XN2=N2
    XN3=N3
    XN4=N4
    XN5=N5
    XN6=N6
    RANF=XN1*T1+XN2*T2+XN3*T3+XN4*T4+XN5*T5+XN6*T6
RETURN
END

```

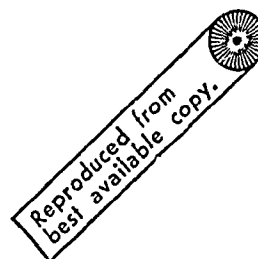
The coding of Table A-6 can be optimized as much as desired so long as the same calculations are performed. This optimization will be a function of the timing considerations for each machine so that there is no point in describing it further here.

The application of the generator involves simply supplying the q-part generator with a q-element array `NS(·)` containing the numbers taken from Table A for the first call with `MODE#0`. Subsequent calls are taken with `MODE=0`.

In order to transform the uniform numbers into gaussian, the coding of Table A-7 can be used. This coding was used for the tests described in this article. The initialization is done the same as for `RANF` through the array `NST(·)` when `MODE#0`. The subroutine generates two numbers every other time and recalls the unused one in between. The returned variable `x` is gaussian with mean 0 and standard deviation 1. Again, if the storage is not protected when the subroutine is not active, it is necessary to save the status words `XR(·)`, `J`, and `TWOPI` in `COMMON`.

Table A-7. FORTRAN-II Coding of Gaussian Generator

```
      FUNCTION GAUSS(NST,MODE)
      DIMENSION NST(1), XR(2)
      IF (MODE) 10,20,10
10    J=2
      TWOPI=8.*ATAN(1)
      XR(1)=RANF(NST,1)
      GO TO 35
20    GO TO (30,40),J
30    J=2
      XR(1)=RANF(NST,0)
35    XR(2)=RANF(NST,0)
      X1=SQRT(ABS(-2.*ALOG(XR(1))))
      X2=TWOPI*XR(2)
      XR(1)=X1*SIN(X2)
      XR(2)=X1*COS(X2)
      GAUSS=XR(1)
      RETURN
40    J=1
      GAUSS=XR(2)
      RETURN
      END
```



## V. Parallel Computational Techniques

### A. Introduction

The success of any numerical technique is directly linked to the application of a suitable computational device for the problem at hand. If the required computational algorithm is completely unstructured then very little can be done in general to improve computation over a straightforward serial machine which performs a single calculation at a time. Most problems contain some independent computational paths, or parallelisms, however, and nearly a decade has been devoted by the computer industry to the problem of identifying and exploiting natural parallelisms of algorithms. The fact that so many approaches to the problem have developed, on the other hand, is testimony to the complexity of the problem in general. In this chapter we will illustrate the fact that Bayes Law calculations are highly structured with essentially arbitrary parallelism, and discuss some of the proposed and existing forms of parallel architecture in relation to this problem.

### B. Parallelism and Bayes Law

Before introducing the alleged parallelism of Bayes law it is appropriate to distinguish among the various forms of parallelism which exist in computers. We observe that the following "levels of computation" exist within a computational system: Electronic, Logic (Bit), Functional Unit, Instruction Stream, Process or Task, Job Step, and Job.

Parallelism at the electronic and bit levels is essentially universal in

**Preceding page blank**

computer architecture today. Some computers have parallelism at the functional unit level (i.e., look-ahead machines). However, very few commercially available machines exhibit parallelism (other than one or two copies) at the instruction stream or higher levels. The reason is clear: hardware duplication is expensive and unless there is a good probability that the extra hardware will always be active (i.e., not idle) then the payoffs rapidly diminish. To gain a quick appreciation for the complexity of the problem consider the evaluation of the simple six term sum  $a+b+c+d+e+f$ : If we only have one adder then we might consider the evaluation in the form  $(((((a+b)+c)+d)+e)+f)$ , (which, of course, is not unique) since this evaluation requires only one temporary storage location. The operation, which takes five add cycles in succession, is diagrammed in Fig. 1.

Suppose, now, you were given two adders which operate independently. Then you might consider the evaluation  $((a+b)+(c+d))+(e+f)$ , which is diagrammed in Fig. 2. Now you have obtained the answer in three add cycles at the expense of two temporary storage locations and one idle adder in the last cycle (since you only need to do five adds). It may be shown that the computation in Fig. 2 is maximally parallel in the sense that no fewer than three add cycles are required, no matter how many adders you have at your disposal. Consider the case with three adders, for example, where after adding  $(a+b)$ ,  $(c+d)$  and  $(e+f)$  simultaneously, you are left in an analogous position to level two in Fig. 2. Thus, after doubling your functional units you have achieved the maximum of 40% decrease in computation time, so that additional units only decrease the profit/cost ratio for this problem. An interesting discussion of the parallel evaluation of arithmetic

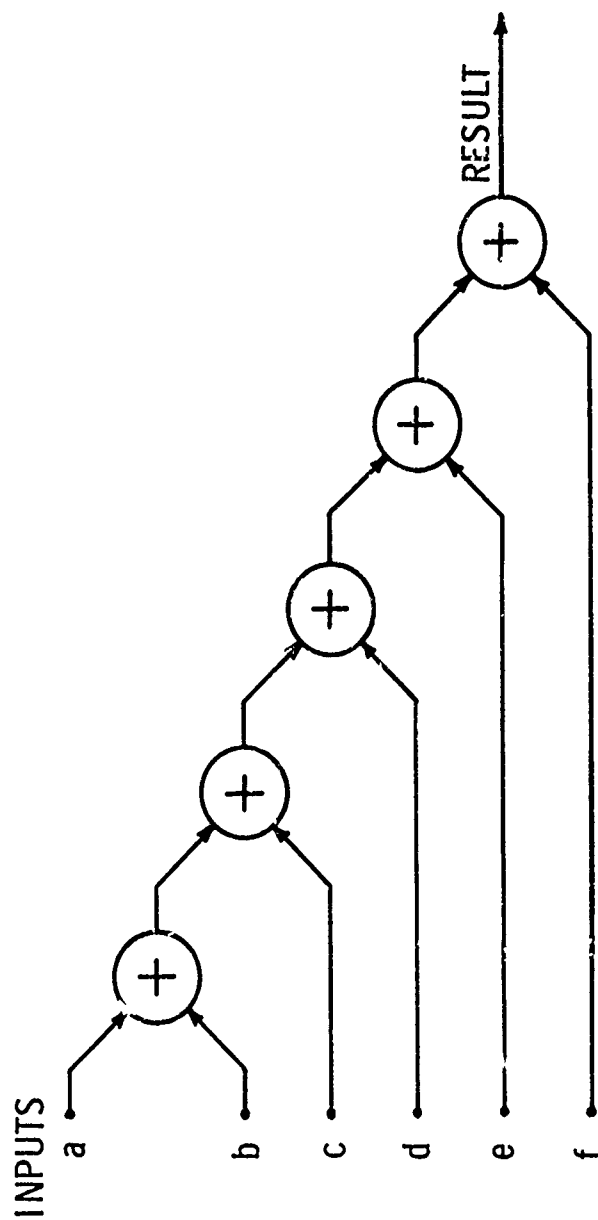


Fig. Serial Evaluation of  $a+b+c+d+f$

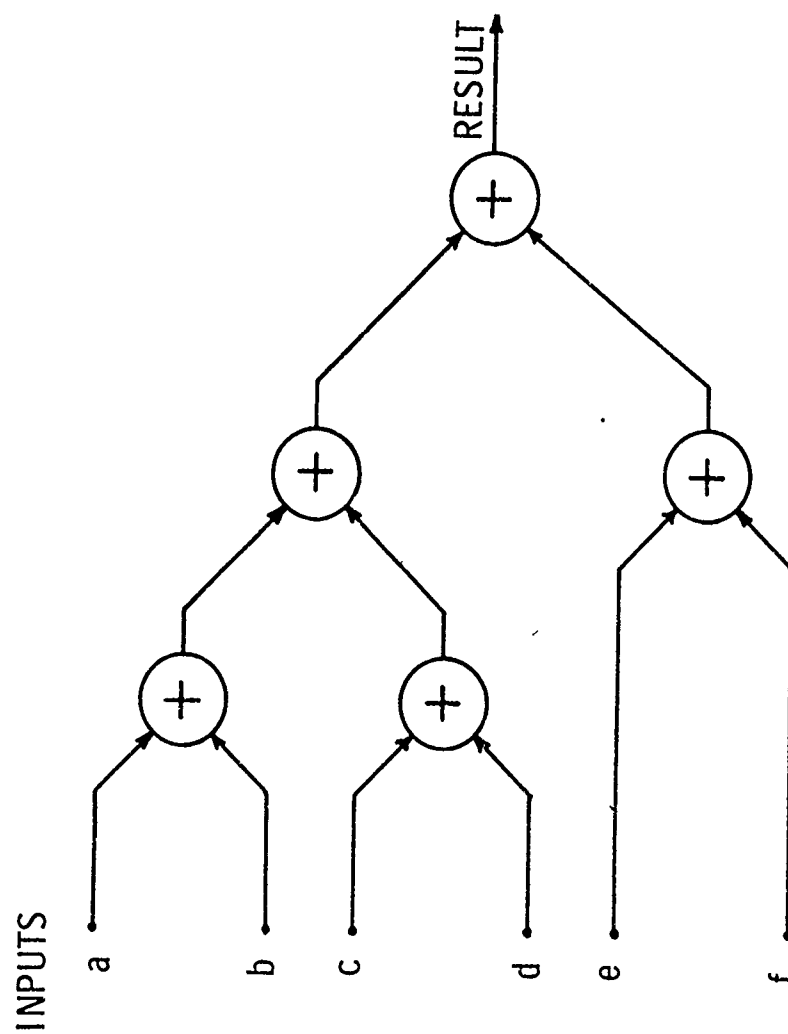


Fig. 2. Maximally Parallel Evaluation of  $a+b+c+d+e+f$

expressions is given by Baer and Bovet [2]. For a definition of maximal parallelism and a development of the associated theory, see Keller [9].

Now that a note of caution has been inserted we proceed to a discussion of highly structured parallel computations, where we use Bayes Law as an example. Bayes Law may be expressed operationally in the form

$$C_{n+1} J_{n+1}(y) = \int d \int T_n(y, x) J_n(x) dx, \quad (1)$$

where  $T_n(y, x)$  is a spatially varying kernel which is a function of the product of probability densities of the noises and the new measurement  $Z_n$ , and  $C_{n+1}$  is the normalizing constant (i.e., the integral of the right-hand side of (1) over all  $y$ ). It should be clear from inspecting (1) that there is implicitly a form of parallelism called for, in that for every  $y$  in  $J_{n+1}$  a  $d$ -dimensional convolution integral must be computed. The explicit form of parallelism will be a function of the nature of the (finite-dimensional) algorithm chosen to implement (1). The point-mass approach of Bucy and Senne [5] maps (1) into an equivalent matrix multiplication. The least-squares series expansions of Hecht [8], Alspach [1], and Center [6], for example, replace (1) by an equivalent update for expansion coefficients. The exact implications on the structure of the computer will of course be dependent upon the form of the algorithm. To be specific, then, since many of the associated problems are similar, we will discuss the matrix multiply analogy to (1) where symbolically we write

$$J'_{n+1}(y_i) = \sum_{j=1}^M T_n(y_i, x_j) J_n(x_j), \quad i = 1, \dots, M, \quad (2)$$

$$C_{n+1} = \sum_{i=1}^M J'_{n+1}(y_i) , \quad (3)$$

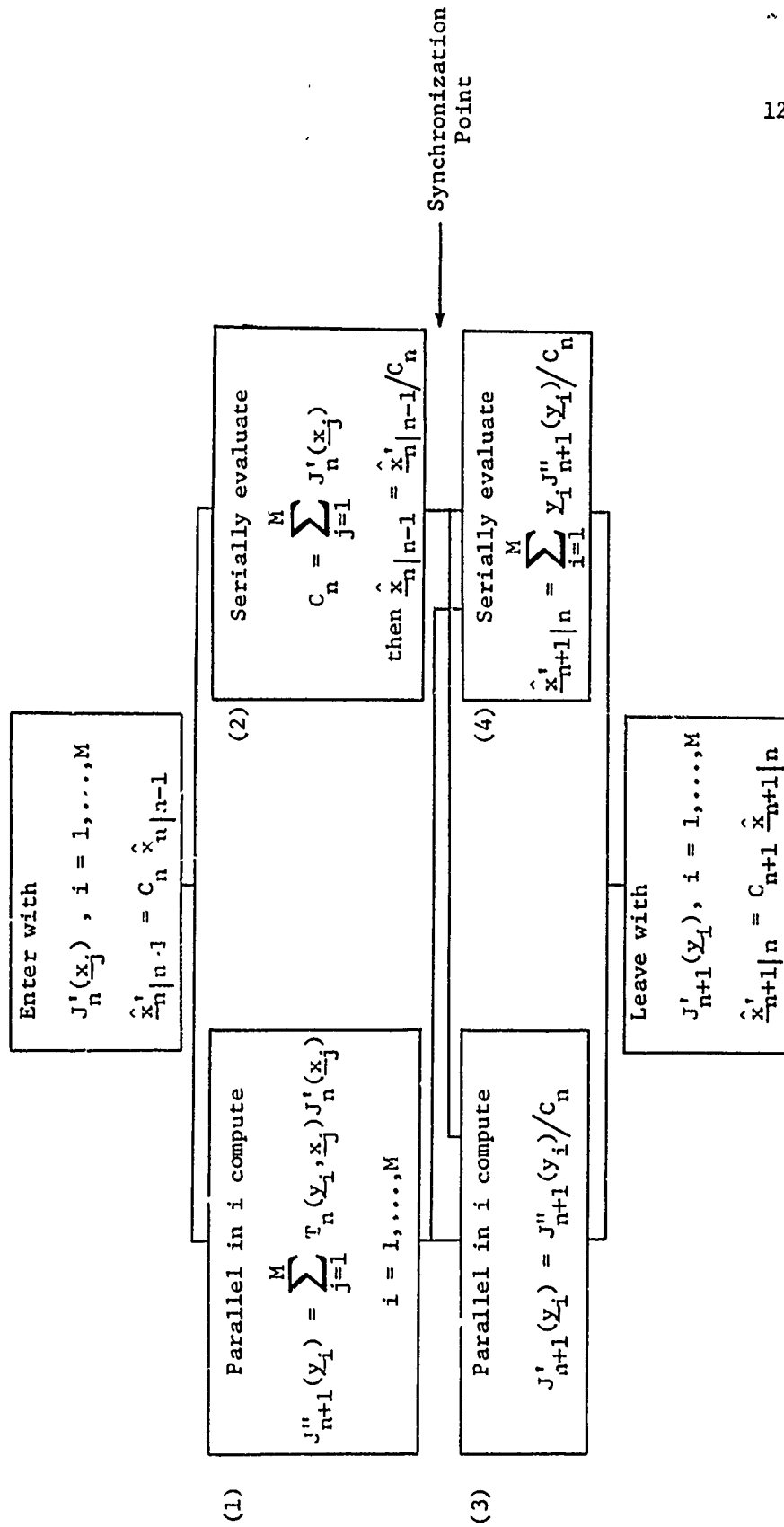
$$J_{n+1}(y_i) = J'_{n+1}(y_i) / C_{n+1} , \quad i = 1, \dots, M , \quad (4)$$

$$\text{and} \quad \hat{x}_{n+1|n} = \sum_{i=1}^M y_i J_{n+1}(y_i) . \quad (5)$$

The overall structure of (2) - (5) (where we choose to discuss the conditional mean estimate as a specific example) illustrate both purely parallel (2), (4) and essentially serial (3), (5) computations. While (by analogy with the example in Figs. 1-2) it is obvious that some parallelism could be built into (3) and (5), it is also clear that a processor which is optimized to perform the totally parallel operations (2) and (4) would be mostly wasted on the scaling and estimation integrals of (3) and (5). Thus it is clear that even in this highly structured problem a combination of computer architectures would be necessary to optimize computational speed with minimum overhead (i.e., idle functional units or processors). Ideally, then, we might envision a parallel computer simultaneously evaluate  $T_n(y_i, x_j) J_n(x_j)$  for  $i = 1, \dots, M$  and accumulates the sum (2) in  $J'_{n+1}(y_i)$  successively for  $j = 1, \dots, M$ . (If we had enough processors and temporary storage we could even imagine tree-structuring the latter computation in the form of Fig. 2.) Imagine, further, that we only had  $J'_n$  to use in (2) so that our  $J'_{n+1}(y_i)$  is off by a factor of  $C_n$ , which, simultaneous to the evaluation of (2), we are evaluating in an auxiliary serial machine, and  $\hat{x}(n|n-1)$ , which we are also evaluating in a serial machine (modulo  $C_n$ ) using  $J'_n$  instead of  $J_n$ . The resulting computation timing could take the form in Fig. 3.

Fig. 3. Combining Serial and Parallel Computations

Assume  $x_i = y_i$ ,  $i=1, \dots, M$



It is reasonable to assume that if the serial and parallel processors in Fig. 3 are on the order of speed, then block 1 would be the slowest, 4 the next slower, 2 the next, and 3 the fastest. Thus, assuming synchronization initially, the serial computer would be idle waiting for block 1 to finish, then block 3 would finish before (4), but since 1 doesn't require the result of 4, the parallel computations may proceed directly without a wait from 3 to 1. Finally, assuming the sum of the execution times of 4 and 2 is less than the sum of the execution times of 1 and 3, the serial processor will again be waiting at the synchronization point in the next cycle when the parallel processor finishes.

Although the computations in Fig. 3 are somewhat simplified (no allowance has been made for floating grids, for example) the figure clearly illustrates the desired tradeoff between essentially serial (overhead) calculations and mostly parallel computations. While it should be understood that a parallel serial computer could be designed around the problem in Fig. 3, it is not reasonable to expect that such a special purpose machine is likely to evolve in the near future. Thus it is more relevant to consider the Bayes estimation problem in light of current existing parallel architectures.

### C. Look-Ahead Processors

The only existing computational device which even closely approximates a "general purpose" parallel processor (with respect to a single instruction stream) is the look-ahead machine, as exemplified by the IBM 360/91 or the CDC 6600 (or 7600). These machines have a multiplicity of distinct (although not necessarily dissimilar) functional

units. The CDC 6600, for example, has two multipliers, two incrementers, an adder, an integer adder, a divider, a boolean unit, a shift unit, and a branch unit for a total of ten distinct units [7]. In addition, there are collections of operand registers (6), resultant registers (2), address registers (8), increment registers (8), and instruction registers (8) to provide for high speed temporary data flow and simultaneous instruction operation together with the above-mentioned hardware, the look-ahead processor has a set of conflict rules whereby the reservations and use of the functional units is directed with the goal of maximum local parallel operation.

Short of assembly language, program there is no guarantee of optimal use of the look-ahead processor hardware. Given the look-ahead rules, there are many situations which will cause the unit to hang-up without all units active. For example, one unit may require an operand to be loaded from the same memory page that a previously implemented store operation has tied up, or a functional unit might be awaiting the second of two operands which is not available for a variety of reasons. There are, however, some programming techniques which will increase the prospects for successful parallelism with the look-ahead processor.\* We will discuss some of these principles in light of the Bayes-Law computation.

Basically one should attempt to reduce the interdependence of adjacent calculations to a minimum. We would start out with the same basic outline as in Fig. 3 and expand the outlines so that locally at

---

\* Although most of the simulations described in this report were performed on a CDC 6600, due to a number of factors no attempt was made to optimize the code.

most one or two arithmetic operations are dependent. This might mean, for example, a partial realization of the operator  $T_n$  including breaking down the exponential function if there is one, performing the partial realizations for all  $j$  in the new array before returning for the next stage in the decomposition of  $T_n$ . The arrays of storage must be laid out with the page partitions of core memory in mind, since sequential access to the same page causes avoidable delays. The CDC 6600, for example, has the page address contained in the least significant address bits, so that adjacent memory addresses are in different memory pages. The obvious implication, then, is that simultaneous operand arrays should be interleaved to maximize the probability that all local computations refer to consecutive core locations.

The fact that look-ahead processing is reasonably successful for most scientific problems results from the fact that most algorithms exhibit some degree of local parallelism, regardless of optimization efforts. When an extremely parallel structure is encountered, however, it is possible to take maximum advantage of look-ahead architecture by maximizing local parallelism, since the look-ahead stack seldom contains more than a few dozen instructions.

#### D. Array Processors

As the name implies, an array processor consists of an array of identical processing elements. The complexity of the individual processors may vary but they all operate on the same instruction stream with different data. As an example, consider the Illiac IV, built in conjunction with University of Illinois by the Burroughs Corporation [3].

The Illiac contains an 8 x 8 array of sophisticated processors, driven in lock step by a Burroughs 6500 computer. Tse and Larson [10], [13] have considered estimation and Bayes calculations on the Illiac structure. Although the Illiac has considerable computational potential, its special structure precludes arbitrary higher level programming. Instead the entire problem computations, array storage, and data manipulations must be customized tailored for this special machine. The necessary synchronization of 64 processors can lead to considerable processor idleness if the computations are not laid out carefully.

#### E. Associative Processors

Another class of parallel machines consists of those containing an associative memory, as typified perhaps by the Goodyear Associative Processor [11]. Each processor word in an associative memory is extremely long (i.e., several hundred bits). Basically the associative memory word may contain several numbers along with some identification bits. Using maximally parallel binary compare logic, certain cells are identified on the basis of their identification bits and then a pre-arranged arithmetic operation or series of operations is performed on the selected words. The concept is certainly an interesting one, but we are unaware of any stand-alone uses other than sorting and cataloguing for the existing machines.

#### F. Pipe-Line Processor

The pipe-line concept takes advantage of the fact that processor and logic speeds are generally much faster than memory speeds, so that

certain compute-bound parallelizable algorithms are constrained to operate at memory speeds in conventional hardware. If large numbers of identical calculations are to be performed on arrays of data, then one possible solution is to page through memory, taking one number from each page and to insert these numbers into a long chain of processors arranged in analogy with a pipeline, with registers between each processing element. Ideally a sequence of computations is performed on each group of data inserted into the line and, if proper allowance is made for queuing-theoretical concerns, the answers are swept off the end register and reinserted into memory at the same average speed that the original numbers were removed. Again, as with all highly structured parallel processors, a considerable amount of program optimization and data structuring is necessary to allow for efficient continuous use of the hardware. The nonavailability of variables to fill the pipe line, as possible, for example, if an isolated (overhead) calculation is required, or if data arrangement operations can't keep up with the pipe-line computations, can be very costly.

In principle, at least, the CDC STAR (the first existing pipe-line machine) promises to be the most nearly ideal structure for the Bayes Law computations in existence. There remains only speculation regarding the ideal computer for Bayes Law, but it wouldn't be surprising if it turned out to be a combination of the machine concepts discussed above in this Chapter. It is expected that actually programmed tests of Bayes Law will provide the best evaluation of the parallelism conjectures we have addressed here. It is clear, however, based on the CDC 6600 experience, that some form of parallelism will be beneficial to the Bayes Law computations.

### G. Hybrid Computer Methods

While most of the present chapter has been devoted to parallel digital computer architectures, it is appropriate to investigate the intrinsic parallel computational structure of the analog computer. Since it is possible to hook up essentially an unlimited number of operational amplifiers operating lock-step in parallel, it appears natural to study the possibility of implementing the essentially parallel aspects of the Bayes-Law computations on an analog device. In his dissertation Miller [12] has, in fact, already undertaken such a study. In the recent paper of Bucy, Merritt, and Miller [4] (reproduced in Additional Appendix F to this report), some of the important details of this research are reported.

Basically, Miller has determined that an efficient use of a hybrid (analog and digital) system is (1) to use the digital computer to compute estimates, normalize densities, and accumulate the Bayes integral computations, and (2) to use the analog computer to compute in parallel the probability density terms (exponentials in the gaussian-noise case), and perform a parallel continuous integration of the unnormalized Bayes integral - stopping at discrete time instances to send the results to the digital computer. While on the surface the speed advantages over conventional digital serial computations look exceedingly attractive, care must be used in evaluating the effects of various analog circuit inaccuracies to determine the attainable accuracy. Miller has used an extensive digital simulator of a hybrid system with programmable

inaccuracies to discover the accumulated effects of realistic analog anomalies such as amplifier noise, offset, drift, nonlinearities, timing mismatch, and limited dynamic range.

Because of the above segmentation of the analog and digital components of the problem, and because Miller is considering point-mass representations of the a posteriori densities in the digital machine, it is natural to consider Miller's hybrid technique an analog generalization of the point-mass approach to the Bayes-Law computations. With that in mind Miller has undertaken comparisons between the two methods on a one-dimensional  $x^3$  sensor problem. In summarizing his conclusions, Miller estimates that in computing one estimate for a four-dimensional problem, one hybrid computer with 250 integrators and multipliers could accomplish the task in 10 seconds, whereas 49 CDC 6600's operating in parallel would require 3 seconds. In summarizing the accuracy, he promises hybrid accuracy of somewhere between 0.1 and 1.0 percent disagreement between the hybrid and the optimal estimates. These conclusions certainly suggest that further development of hybrid systems may indeed be profitable.

## References

- [ 1 ] D.S. Alspach, "A Bayesian Approximation Technique for Estimation and Control of Time Discrete Stochastic Systems," Ph.D. Dissertation, University of California, San Diego, 1970.
- [ 2 ] J.L. Baer and D.P. Bovet, "Compilation of Arithmetic Expressions for Parallel Computations," Information Processing 1968, North-Holland Pub. Co., Amsterdam, 1969, 340-346.
- [ 3 ] W.J. Bouknight, et al. "The Illiac-IV Systems," Proc. IEEE, 60 (1972), 369-388.
- [ 4 ] R.S. Bucy, M.J. Merritt, and D.S. Miller, "Hybrid Computer Synthesis of Optimal Discrete Nonlinear Filters," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, 1971, 59-87.
- [ 5 ] R.S. Bucy and K.D. Senne, "Digital Synthesis of Non-Linear Filters," Automatica, 7 (1971), 287-298.
- [ 6 ] J.L. Center, Jr., "Practical Nonlinear Filtering of Discrete Observations by Generalized Least Squares Approximation of the Conditional Probability Distribution," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, 1971, 88-99.
- [ 7 ] Control Data 6400/6500/6600 Computer Systems Reference Manual, CDC Publication 60100000, St. Paul, 1968, Chaps 2 and 3.
- [ 8 ] C. Hecht, "Digital Realization of Non-Linear Filters," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, 1971, 152-158.
- [ 9 ] R.M. Keller, "On Maximally Parallel Schemata," IEEE Conf. Rec. 11th Annual Symp. on Switching Theory and Automata Theory, 1970, 32-50.
- [10] R.E. Larson and E. Tse, "Modal Estimation and Parallel Computers," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, 1971, 188-197.
- [11] W.C. Meilander, "The Associative Processor in Aircraft Collision Prediction," NAECON Proc., 1968.
- [12] D.S. Miller, "Hybrid Synthesis of Optimal Discrete Nonlinear Filters," Ph.D. Dissertation, University of Southern California, 1971.
- [13] E. Tse, "Parallel Computation of the Conditional Mean State Estimate for Nonlinear Systems," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, 1971, 385-394.

## VI. A Passive Receiver: Bearings - Only Tracking (AWACS)

### A. Introduction

The first nonscalar example of the general Bayesian nonlinear estimation problem which has been studied in considerable detail was introduced by Bucy, Geesey, and Senne [3]. The problem involves the detection and tracking of targets based on passive (i.e., without the use of radar or sonar) receivers. The basic problem description has arisen from a variety of applications, most of which involve some form of strategic sleuthing, either by an airborne, a landbased, or a sea-going observer. In these applications it is presupposed that the only form of measurement available to the sensor is bearing information, resulting from directed energy radiated from the target in the form of radar or sonar. Thus, in order to extract range information from the noisy bearing measurements it is necessary for either multiple sensors to be employed at a variety of known geographic positions, or a single sensor to be moved through an appropriate sequence of positions (i.e., a known orbit). The latter sensor is most important for a variety of reasons: (1) there is no need for multiple sensor location calibration, a problem which is frequently as difficult as the passive receiver problem, (2) it is frequently impractical to devote a multiplicity of sensors to track only one target, and (3) it is usually desirable to have the sensor and its support vehicle be one self-contained and

**Preceding page blank**

independent unit. Thus a sequential estimation problem is identified with the passive-receiver.

To be specific we will assume that the target is a radar-bearing, slowly-moving object, such as a surveillance ship, and we will let the sensor-bearer be an orbiting aircraft, such as the Airborne Warning And Command System (AWACS), which is currently under Air Force contractor development. We don't intend to describe the complete AWACS problem in detail here, but only to appeal to one of its primary objectives - passive tracking, as an example of the kind of problem we are studying. We expect that the very simplified example we have chosen to discuss here will be sufficiently interesting so as to motivate a method of attack for many other aspects of the AWACS problem.

The most important characteristic of the bearings-only receiver problem is shown in Fig. 1. Here we see the basic geometry of the receiver (taken to be restricted to two-dimensions for simplicity). The target  $T$  is located at coordinates  $(x_1, x_2)$  and the sensor  $S$  is located at coordinates  $(S_1, S_2)$ . We assume that sensor is capable of measuring the bearing angle  $h(\underline{x})$ , given by

$$h(\underline{x}) = \tan^{-1} \left[ \frac{x_2 - S_2}{x_1 - S_1} \right]. \quad (1)$$

Of course the measurement of  $h(\underline{x})$  will in most cases be contaminated by noise, which we will assume to be additive, resulting in the measurement scheme

$$dz(t) = h(\underline{x})dt + dv(t) \quad (2)$$

where  $h(\underline{x})$  is given by (1). Some possible sources of the noise might be receiver-noise (i.e., shot noise), Electronic Counter Measures (ECM),

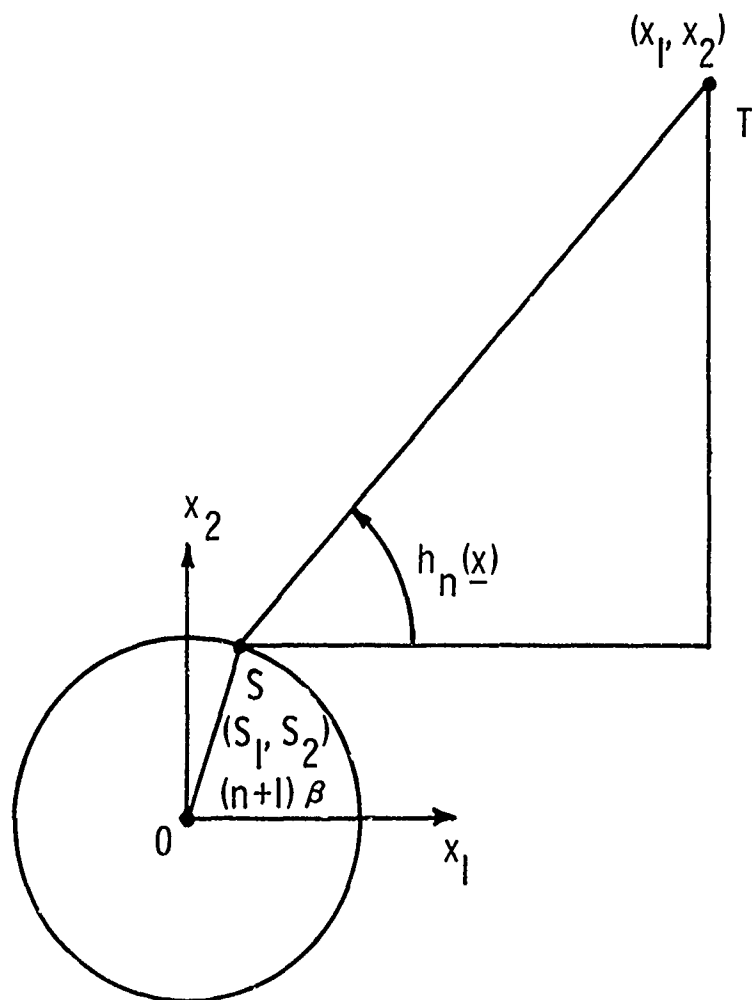


Fig. 1. Typical Passive Receiver Geometry

or atmospheric noise. In any case we will assume that the noise is characterized by a known distribution (specifically taken to be Gaussian for purposes of example).

As was mentioned above, the sensor using the bearings-only scheme (1) would not be able to measure the range  $R = [(x_1 - S_1)^2 + (x_2 - S_2)^2]^{1/2}$ , regardless of the target motion or the number of measurements. That is, the range is unobservable to a stationary sensor. Thus, it is necessary to require that the sensor proceed in a prescribed orbit, which we will take to be a unit circle, for example. We further assume that the measurements are to be taken only at discrete instants in time (at, say,  $t = n\Delta$ ,  $n = 0, 1, 2, \dots$ ). The resulting measurement scheme takes the form of

$$z(n) = \tan^{-1} \left[ \frac{x_2(n) - \sin \beta (n+1)}{x_1(n) - \cos \beta (n+1)} \right] + v(n) \quad (3)$$

where we have assumed that the sensor proceeds counter-clockwise around the circle at the rate  $\beta$  radians per sample.

In order to make the bearings-only problem into a non-trivial filtering problem, it is necessary to introduce a stochastic model for the allowable motions of the target. There is, of course, a wide variety of useful possibilities which might be selected. Consider, for example, the cases of essentially stationary (but with random velocity vector), essentially constant velocity (but with random accelerations), and the various orders of acceleration models. An example of an essentially stationary target would be an unanchored, but undriven ship, being buffeted by the waves. An essentially constant velocity would result possibly from an "airliner target," or one that is attempting to

navigate a fixed course, but is subject to random aerodynamic forces. Finally, the various orders of acceleration models would usually result from maneuvering targets. The dimension of the underlying state space increases by two for each increase in degree of freedom of the motion model, so that a stationary target is described in two dimensions, a constant velocity model has four, and a constant acceleration model has six. If the acceleration is assumed to be dynamic, then the number of control states may increase indefinitely. In addition, one may envision measurement bias states, as well as many other unknown parameters, so that an initially modest problem turns into a monster.

We could easily have chosen the most complicated form of the problem for this example. On the other hand we feel that is counter-productive for two reasons: (1) our principle goal is to illustrate, with very simple examples, what the Bayes-law formulation says about nonlinear estimators, and hopefully to provide a basic understanding of the characteristics of optimal estimates, and (2) we are obliged to perform extensive Monte Carlo experiments in order to learn what characteristics of the conditional densities are the most important for estimation purposes, and to provide high-confidence performance analyses. Consequently, we have chosen to study the simplest example of the passive receiver problem which still has dynamics - the essentially stationary target (with random velocity). The equations of the assumed model take the form

$$\underline{x}(n+1) = \phi \underline{x}(n) + \underline{u}(n) , \quad (4)$$

where  $\phi$  is a  $2 \times 2$  transition matrix, and  $\{\underline{u}(n)\}$  is a sequence of mutually independent random vectors, having covariance  $Q$ . If the  $\phi$

matrix is an identity matrix, then  $\underline{x}(n)$  performs a random walk in two dimensions. We will also be interested in studying examples of stable as well as unstable transitions.

### B. The Linearized Estimator

One obvious candidate for an approximate estimate of the position of the target is the linearized or extended Kalman-Bucy filter. If we choose to linearize about the current filter estimate  $\hat{\underline{x}}_2(n|n)$ , then the linearized measurement function  $H'(n)$  takes the form

$$\begin{aligned} H(n) &= \nabla_{\underline{x}} h[\hat{\underline{x}}_2(n|n)] \\ &= \left[ \frac{-b}{a^2 + b^2} \frac{a}{a^2 + b^2} \right], \end{aligned} \quad (5)$$

where we have taken  $a \triangleq \hat{x}_2^1(n|n) = \cos \beta(n+1)$ ,  $b \triangleq \hat{x}_2^2(n|n) = \sin \beta(n+1)$ .

Formally we write the equations for the linearized filter as

$$\hat{\underline{x}}_2(n+1|n) = \Phi \hat{\underline{x}}_2(n|n), \quad \hat{\underline{x}}_2(0|-1) = \underline{0}, \quad (6)$$

$$\hat{\underline{x}}_2(n+1|n+1) = \hat{\underline{x}}_2(n+1|n) + A(n+1) \left\{ z(n+1) - h_{n+1}[\hat{\underline{x}}_2(n+1|n)] \right\}, \quad (7)$$

$$A(n+1) = P(n+1|n)H(n+1)'R^{-1}[H(n+1)P(n+1|n)H(n+1)'+R]^{-1}, \quad (8)$$

$$P(n+1|n+1) = [I - A(n+1)H(n+1)]P(n+1|n)[I - A(n+1)H(n+1)]' + A(n+1)R A'(n+1), \quad (9)$$

$$P(n+1|n) = \Phi P(n|n)\Phi' + Q, \quad (10)$$

and  $P(0|-1) = \Gamma(0)$ . In attempting to apply (6) - (10), however, a difficulty arises in the timing requirements, since the filter update (7) requires  $A(n+1)$  which in turn requires  $H(n+1)$  in (8) and (9), but  $H(n+1)$  depends on  $\hat{\underline{x}}_2(n+1|n+1)$ , which is not yet available!

A standard solution to the dilemma is to evaluate the gradient (5) at the predicted estimate  $\hat{x}_p(n|n-1)$ . If the predicted and filtered estimates differ substantially, however, or if the measurement function is sensitive, such as the case in (5), then it is possible to partially restore the proper timing by iterating (5), (9), (8), and (7), starting with (5) evaluated at  $\hat{x}_p(n|n-1)$  as above and continuing until the filtered estimate (7) stabilizes. After each iteration the current "quasi-filtered" estimate from (7) is used to evaluate the next gradient (5). After the filtered estimate converges the prediction sequence (10) and (6) can then be performed, thereby completing the estimation cycle. Denham and Pines [7] have reported substantial improvement by using the iterated filter for some tracking problems.

It turns out that if the linearized filter or iterated filter fails to perform properly, there are essentially an unlimited number of "fixes" or engineering modifications to be found in the literature. Most fixes involve some form of dynamic stabilization of the equations (6) - (10) to avoid such unpleasant characteristics as divergence of the errors, negative definite P's, zero gains (A), or unstable estimates. We find, in fact, that a simple limiting operation removes the tendency for this particular linearized filter to go unstable (see Appendix A) for certain initial conditions.

### C. Application of Nonlinear Filtering

The two approaches which we have applied to the Bayes-Law calculations for the passive tracking problem include point masses with a floating grid and Gaussian sums. De Figuierido [6] has recently

applied spline functions to the problem, but we have not had opportunity to confirm his results.

In applying the point mass technique we endeavored to utilize the full generality of the method as described by Bucy and Senne [4] (see Chapter III). That is, we centered the grid predictively (including rotation to align the grid coordinates with the eigenvectors of the predicted covariance matrix). We also took advantage of the unimodal character of most of the densities and used the ellipsoidal tracking scheme for calculating the Bayes integral. The equations implemented, in fact, were so similar to the ideal that there is little point in repeating them here.

In his dissertation, Alspach [1] observed that the gaussian sum approximation to the filtering problem could be simplified whenever the plant was linear, the noise is gaussian and the measurement nonlinearity has certain symmetry properties. The simplification involves identification of region B in the state space where  $\underline{z}(n) = \underline{h}(\underline{x})$ . Then, if B is a simple surface in the state space, the gaussian sum may be concentrated in B. It turns out that in the passive receiver problem, the region B is a simple hyperplane passing through the receiver position. This may be demonstrated as follows: Let

$$z = \tan^{-1} \left( \frac{x_2 - s_2}{x_1 - s_1} \right),$$

then set

$$B = \left\{ (x_1, x_2) : z = h(x_1, x_2) \right\},$$

or, equivalently,

$$\tan z = \frac{x_2 - S_2}{x_1 - S_1} ,$$

resulting in

$$x_2 = (\tan z)x_1 + S_2 - S_1 \tan z . \quad (11)$$

Furthermore, simple geometric observations allow one to eliminate all points outside or inside the sensor orbit, depending on the apriori information concerning the target's position. Thus the sensor density function is symmetric about the line (11) and cone shaped (see Alspach [1] for figures). To implement the gaussian sum approximation, then, Alspach simply places gaussians on the line (11), resulting in a one-dimensional realization of the problem.

## References

- [1] D.L. Alspach, "A Bayesian Approximation Technique for Estimation and Control of Time-Discrete Stochastic Systems," Ph.D. Dissertation, University of California, San Diego, 1970.
- [2] D.L. Alspach and H.W. Sorenson, "Approximation of Density Functions by a Sum of Gaussians for Nonlinear Bayesian Estimation," Proc. Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1970, 19-31.
- [3] R.S. Bucy, R.A. Geesey, and K.D. Senne, "Passive Receiver Design via Nonlinear Filtering Theory," Proc. Third Hawaii International Conf. on System Sciences, Vol I, 1970, 477-480.
- [4] R.S. Bucy and K.D. Senne, "Digital Synthesis of Nonlinear Filters," Automatica 7 (1971), 287-298.
- [5] R.S. Bucy and K.D. Senne, "A Two-Dimensional Passive Ranging Experiment using Optimal Nonlinear Filtering," Air Force Weapons Laboratories Computer Films No. 71-0330-02, March 1971.
- [6] R.J.P. deFigueiredo and Y.G. Jan, "Spline Filters," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1971, 88-99.
- [7] W.F. Denham and S. Pines, "Sequential Estimation when Measurement Function Nonlinearity is Comparable to Measurement Error," AIAA J. 4 (1966) 1071-1076.
- [8] K.D. Senne, "Computer Experiments with Nonlinear Filters," Proc. Second Symp. on Nonlinear Estimation and Its Applications, San Diego, 1971, 314-324 (Misprinted - see Additional Appendices of this report for corrected version).

## Appendix A. First Monte Carlo Experiments

### Point Masses versus Linearized

In this appendix we summarize the initial experiments which we performed on the Burroughs B5500 machine and reported in Bucy and Senne [4]. We make some revised accuracy estimates concerning the reliability of the experiments and some additional interpretations of the results.

The parameters for the initial tests were selected essentially arbitrarily to be

$$\Phi = \begin{bmatrix} 0.5 & 0 \\ 0 & 1 \end{bmatrix}, \quad Q = \begin{bmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{bmatrix},$$

$$\beta = 1, \quad R = 0.1$$

and  $J_0 \sim N(\underline{0}, I).$

In addition we stored 4 standard deviations on the floating grids (based on the predicted covariance) and used a 3 standard-deviation ellipse for integrating Bayes law.

The raw Monte Carlo data may be seen in Table A-1, taken from Bucy and Senne [4]. The estimates for one-step prediction  $\hat{\underline{x}}(n|n-1)$  were computed for  $n=1, \dots, 10$ , with  $\hat{\underline{x}}(0|-1) = \underline{0}$  as an initial condition. In addition, the ideal pure prediction covariance  $\Sigma_n = \Phi \Sigma_{n-1} \Phi' + Q$  is tabulated for reference. The pure prediction covariance is the theoretical error performance of the estimate  $\hat{\underline{x}}_2(n+1) = \Phi \hat{\underline{x}}_2(n)$ ,  $\hat{\underline{x}}_2(0) = \underline{0}$ , or, equivalently, the "zero-state" or "no-information" optimal estimate. The zero-state predictor covariance may be interpreted

as the upper bound on the performance of the optimal nonlinear predictor.

The error performance of the previous filtered estimate corresponding to the entries in Table 2 may be found by inverting the transformation

$P(n+1|n) = \Phi P(n|n)\Phi' + Q$ , since the dynamics are linear. The results

are shown in Table A-2, also taken from Bucy and Senne [4].

Table A-1. Monte Carlo Performance  
of the Optimal and Linearized Predictors

n	Optimal nonlinear predictor			Linearized predictor			Zero state predictor	
	Average error	Average covariance		Average error	Average covariance		Theoretical covariance	
1	-0.094	0.359	0.230	-0.371	0.645	-0.765	0.350	0.050
	0.078	0.230	0.857	0.477	-0.765	2.728	0.050	1.100
2	-0.008	0.146	0.068	-0.277	0.401	-0.183	0.188	0.075
	0.042	0.068	0.599	0.081	-0.183	1.255	0.075	1.200
3	-0.004	0.131	0.051	-0.158	0.222	0.012	0.147	0.088
	0.026	0.051	0.354	-0.041	0.012	0.623	0.088	1.300
4	0.007	0.127	0.088	-0.081	0.158	0.074	0.137	0.094
	0.003	0.088	0.351	-0.084	0.074	0.449	0.094	1.400
5	0.015	0.129	0.069	-0.069	0.155	0.113	0.134	0.097
	-0.022	0.069	0.342	-0.106	0.113	0.723	0.097	1.500
6	0.038	0.177	0.099	-0.128	0.255	0.589	0.134	0.098
	0.025	0.099	0.369	-0.315	0.589	3.813	0.098	1.600
7	0.075	0.217	0.140	-0.188	0.291	0.824	0.133	0.099
	0.065	0.140	0.416	-0.362	0.824	6.439	0.099	1.700
8	0.063	0.159	0.112	-0.186	0.223	0.549	0.133	0.100
	0.093	0.112	0.455	-0.316	0.549	5.890	0.100	1.800
9	0.007	0.130	0.076	-0.141	0.164	0.262	0.133	0.100
	0.043	0.076	0.364	-0.321	0.262	4.950	0.100	1.900
10	0.026	0.136	0.081	-0.146	0.164	0.148	0.133	0.100
	0.036	0.081	0.326	-0.428	0.148	3.978	0.100	2.000

Table A-2. Monte Carlo Performance of Optimal  
and Linearized Filters

n	Optimal nonlinear filter		Linearized filter	
	Average covariance		Average covariance	
1	1.038	0.361	2.180	-1.630
	0.361	0.757	-1.630	2.628
2	0.184	0.036	1.206	-0.465
	0.036	0.499	-0.465	1.155
3	0.123	0.003	0.487	-0.077
	0.003	0.254	-0.077	0.523
4	0.108	0.075	0.232	0.049
	0.075	0.251	0.049	0.349
5	0.117	0.039	0.221	0.126
	0.039	0.242	0.126	0.623
6	0.309	0.097	0.620	1.078
	0.097	0.269	1.078	3.712
7	0.469	0.180	0.764	1.547
	0.180	0.316	1.547	6.339
8	0.234	0.124	0.491	0.999
	0.124	0.355	0.999	5.790
9	0.119	0.053	0.257	0.425
	0.053	0.264	0.425	4.850
10	0.142	0.063	0.256	0.196
	0.063	0.226	0.196	3.878

In reviewing these experimental results it must be pointed out that the confidence analysis of Bucy and Senne [4] is in error. The one-sigma gaussian confidence level is  $(2\sigma^4/N)^{1/2}$ , and not  $(3\sigma^2/N)^{1/2}$ , as given in the original paper (see Chapter IV). Using the results of Chapter IV we will now assess the accuracy of the above test results. First we convert the diagonal terms of the sampled covariances into their three

standard deviation confidence limits, taken from Chapter IV. The results are given in Table A-3 for the predictor covariances of Table A-1. The nonlinear predictor results were based on  $N=500$  Monte Carlos and the linearized predictor was run for  $N=2000$ .

Table A-3. Monte Carlo  
Confidence Intervals for Predictors

n	Optimal Nonlinear Predictor		Linearized Predictor	
	First Coordinate	Second Coordinate	First Coordinate	Second Coordinate
1	0.302→0.443	0.720→1.058	0.589→0.713	2.492→3.014
2	0.123→0.180	0.503→0.739	0.366→0.443	1.146→1.387
3	0.110→0.162	0.298→0.437	0.203→0.245	0.569→0.688
4	0.107→0.157	0.295→0.433	0.144→0.175	0.410→0.496
5	0.108→0.159	0.287→0.422	0.142→0.171	0.660→0.799
6	0.149→0.218	0.310→0.455	0.233→0.282	3.483→4.213
7	0.182→0.268	0.350→0.513	0.266→0.322	5.881→7.114
8	0.137→0.196	0.382→0.562	0.204→0.246	5.380→6.507
9	0.109→0.160	0.306→0.449	0.150→0.181	4.521→5.469
10	0.114→0.168	0.274→0.402	0.150→0.181	3.633→4.395

Almost immediately we ascertain from comparing Table A-3 with Table A-1 that the "optimal" estimate can not possibly be optimal for iterations 6, 7, and 8, since the entire  $3\sigma$  confidence band lies above the zero-state result for those samples. In fact we observe a periodicity in the errors for both estimators with period approximately equal to the rotational period of the sensor (between 6 and 7 samples).

We could also have made a study of the zero-bias reliability of the estimates. But, owing to the dubious value of the data, we choose to proceed on to the more recent experiments.

Appendix B. More Recent Experimental Results:  
Point Masses versus Gaussian Sums

The early experiments reported in the previous appendix raised a lot of questions which demanded more tests to resolve. What made the errors cyclic modulo  $2\pi$ ? Why was the linearized filter unstable? Alspach and Sorenson [2] revealed another "linearized" filter which did not have instabilities. Thus it was imperative that we closely compare our results with theirs in order to explain this anomolous discrepancy.

Using an estimate comparison test of the two linearized filters it was discovered that Alspach and Sorenson had modified the estimate update equation to the form

$$\hat{\underline{x}}_1(n+1|n+1) = \hat{\underline{x}}_1(n+1|n) + A(n+1) \left\{ [z(n+1) - h[\hat{\underline{x}}_1(n+1|n)] + \pi] \bmod 2\pi - \pi \right\}, \quad (B-1)$$

instead of the original form (7). We may intuitively rationalize the success of the modified measurement scheme by examining Fig. B-1. Suppose the working range of the sensor is  $[-\pi, \pi)$  with zero referring to the first coordinate axis. The figure shows two typical situations which might arise if the target is allowed to be inside the sensor orbit, as is the likely case if  $J_0 \sim N(0, I)$ . Case 1 represents the situation for  $\beta(n+1) \approx 0$ . In this case  $h(\underline{x}_1)$ , the true bearing, is positive, while the bearing to the estimated position  $\hat{\underline{x}}_1$  is negative. The result is a difference in bearing greater than  $\pi$ . On the other hand, if both the estimate and the target remain inside the sensor orbit, then some nonzero  $\beta(n+1)$  (such as given in Case 2 of the figure) would result in bearing and estimated bearing of the same sign, so that

**Preceding page blank**

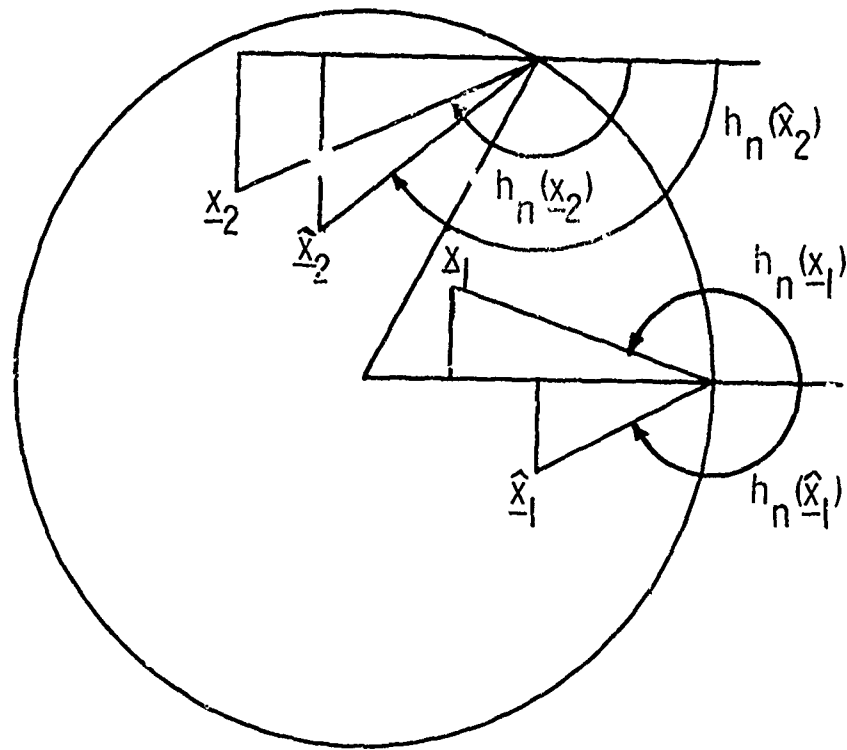


Fig. B-1. Typical Geometry of the "Old Problem"  
 Illustrates Periodicity of Errors

the difference is always less than  $\pi$ . Thus the figure accounts for the modularity of the error performance discovered in the early experiments.

Now the filter update equation (7) contains the term  $z - h(\hat{x})$ , which in turn may be expressed as  $h(x) - h(\hat{x}) + v$ , so that even if  $h(x) - h(\hat{x})$  has an effective range from  $-\pi$  to  $\pi$ , the additive noise is gaussian and may take on all values. Thus, if the noise magnitude is usually small compared with the magnitude of  $h(x) - h(\hat{x})$ , it would be expected that the modulation of the difference  $z - h(\hat{x})$  as in (B-1) would lead to improved performance. The "fixed" linearized performance is in fact dramatically improved, as illustrated in Table B-1, which shows the Monte Carlo results for the "old problem".

Table B-1. Monte Carlo Averaged Sum Squared Error  
Performance for Predictors - Old Problem

Sample	Mean-State Predictor	Iterated Linearized Predictor	"Fixed" Linearized Predictor	Gaussian Sum Predictor	Floating Grid Predictor
1	1.250	4.334	1.468	0.757	0.866
2	1.388	1.678	0.974	0.596	0.674
3	1.447	0.651	0.562	0.457	0.631
4	1.537	0.535	0.550	0.473	0.408
5	1.634	0.828	0.617	0.518	0.403
6	1.734	3.187	0.169	0.531	0.464
7	1.833	4.927	0.540	0.471	0.674
8	1.933	4.912	0.618	0.553	0.454
9	2.03	4.248	0.564	0.578	0.377
10	2.133	3.763	0.619	0.641	0.443
	Unstable	Unstable	Stable	Stable	Stable

What is shown the table is the trace of the sampled covariance matrices for the Mean-State (ideal predictor), the Iterated-Linearized, the Fixed-Linearized, the Gaussian Sum, and the Floating Grid Predictors. Table B-1 is taken from Senne [8], which was printed incorrectly due to an editing error. The fixed-linearized predictor is seen in the table to be stable, and, based on only 100 Monte Carlos, essentially equivalent to both the Gaussian-Sum and Floating-Grid predictors. It turns out, in fact, that the "old problem," as depicted in Fig. B-1, is far easier than originally intended. The sensor is given considerable new information with each sample if it orbits around the target at the high rate of  $\beta=1$  radian per sample. Thus it is not surprising to find that it is relatively easy to design an approximate estimator which performs very close to optimum. The straightforward linearized predictor performs miserably, it happens, as discussed above. We can conclude, however, that the old problem is not sufficiently difficult to demonstrate the difference in performance between the various estimators.

We return thus to the originally intended geometry, illustrated in Fig. B-2, where the initial density  $J_0 \sim N(\begin{bmatrix} 3 \\ 3 \end{bmatrix}, I)$ ,  $R=0.01$ , and  $F=I$ , so that the target is undergoing pure random walk in both dimensions. The Monte Carlo performance summary for the "new problem" is given in Table B-2, where we may discern that the linearized predictor (in this case the "fix" is unnecessary) converges more slowly to steady state than the optimal estimates, but the accuracy of the Monte Carlo (100 sample paths) still precludes discrimination for steady-state operation. Thus we have come to the conclusion that the passive receiver with gaussian noises is very linearizable in that the conditional densities are very nearly gaussian, at least in range-bearing coordinates.

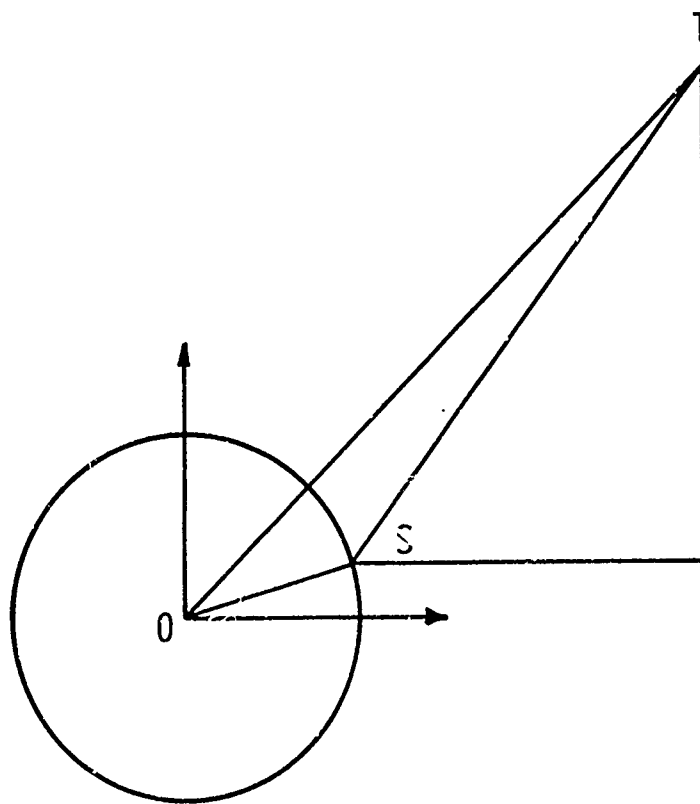


Fig. B-2. "New Problem" Geometry without  
Periodic Errors

Table B-2. Monte Carlo Averaged Sum-Squared Error  
Performance for Predictors - New Problem

Sample	Mean-State Predictor	Linearized Predictor	Gaussian Sum Predictor	Floating Grid Predictor
1	2.200	1.665	11.436*	0.955
2	2.400	1.928	1.723	1.020
3	2.600	2.187	1.495	1.006
4	2.800	2.229	1.321	1.083
5	3.000	2.190	1.208	1.118
6	3.200	2.145	1.456	1.359
7	3.400	1.830	1.625	1.521
8	3.600	2.105	1.437	1.370
9	3.800	2.136	1.423	1.132
10	4.000	2.349	1.286	1.316
11	4.200	2.370	1.431	1.493
12	4.400	2.114	1.533	1.515
13	4.600	1.883	1.475	1.345
14	4.800	1.752	1.398	1.268
15	5.000	1.752	1.553	1.441
16	5.200	1.850	1.617	1.534
17	5.400	1.627	1.275	1.222
18	5.600	1.561	1.271	1.238
19	5.800	1.620	1.630	1.314
20	6.000	1.688	1.803	1.556

We note in passing that the first sample estimate of the Gaussian sum predictor suffers from a transient at about Monte Carlo number 50,

so that the number in the table is inaccurate. The filter recovers stability, however, so that this number may be ignored.

### Appendix C. A Movie of Conditional Densities

In the previous two appendices the relative success of the linearized-type predictor can be related to the validity of the Gaussian approximation to the conditional density functions. The simplest way to destroy the validity of the Gaussian assumption is to provide a nongaussian initial density function. Consider, for example, the detector geometry illustrated in Fig. C-1. If there were a sequence of known reflecting ionospheric layers above the aircraft observer and we were given an apriori distribution on the transmitter's power, then it is conceivable that we might want to integrate over all elevations to maximize detectability, thus introducing a multimodal range ambiguity as shown in the figure. Probabilistically, the initial condition would be obtained by taking the product of the multimodal range ambiguity density with the bearing ambiguity density. The result might look as in Fig. C-2, where no particular scale is intended. As soon as the aircraft's detector circuits obtain a reliable detection, the aircraft banks left into a circle of unit radius and activates a high sensitivity receiver which is tuned to one elevation.

In order to study the evolution of the conditional densities a movie was made by choosing the parameters

$$F = \begin{bmatrix} 1.000 & 0.000 \\ 0.000 & 1.001 \end{bmatrix}, \quad Q = \begin{bmatrix} 0.050 & 0.025 \\ 0.025 & 0.050 \end{bmatrix},$$

$$R = 0.01, \quad \beta = 0.01, \quad \text{and}$$

$J_0$  was chosen as the sum of four Gaussian densities with one cross range

**Preceding page blank**

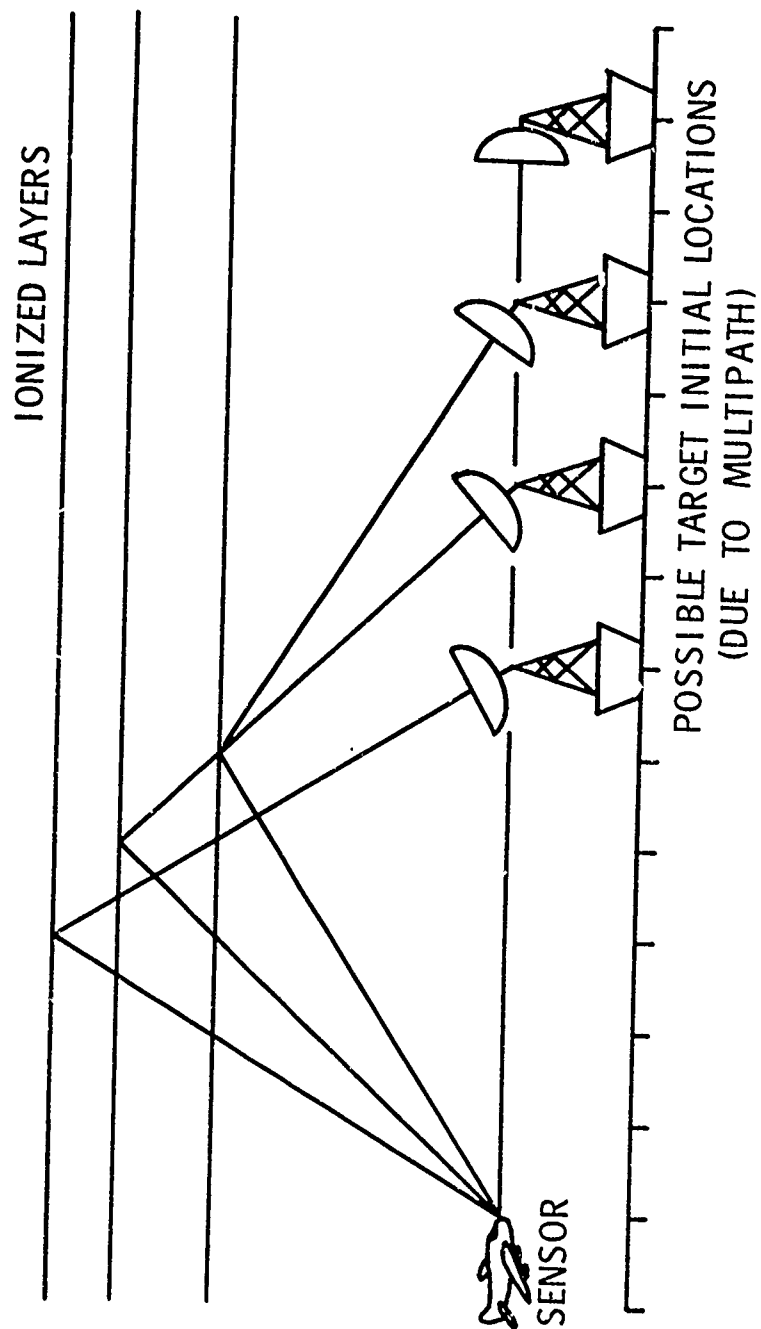


Fig. C-1. Detection Geometry in the Presence of  
Multipath Reception

standard deviation corresponding to 0.25 radian in bearing from the sensor and a down-range standard deviation of 0.25 in each mode (corresponding to the unknown transmitter power). If the densities are plotted isometrically as viewed from the direction shown in Fig. C-2, then some of the more interesting examples from the movie [5] are shown in Fig. C-3 with sample numbers as given. As seen in Fig. C-3, the density first peaks up on the wrong range when very little information is available from the measurements, and later makes a dramatic change to the correct mode and stabilizes. This behavior is clearly illustrated in the solid plot of Fig. C-4 of the resulting prediction error magnitude. In contrast the dotted curve in Fig. C-4 shows the performance for the same sample path of the linearized predictor, which is almost identical to the mean state estimate for the problem. Fig. C-4 illustrates two interesting phenomena. The linearized approximation can be made useless by only a large uncertainty in initial conditions. On the other hand, the nonlinear estimator, which contains the more detailed information about the conditional densities may still perform satisfactorily. In addition, the optimum system creates a certain amount of optimism for itself initially and calculates a relatively high-confidence incorrect estimate. Experience with several such sample paths has resulted in the conclusion that the situation in Fig. C-4 is just about the worst possible case, that on the average the performance is without much improvement for about half the iterations (0.5 radian of sensor rotations), and then proceeds to converge at about the same rate as the previous examples. The linearized predictor, on the other hand, without a suitable initial condition doesn't ever converge at all.

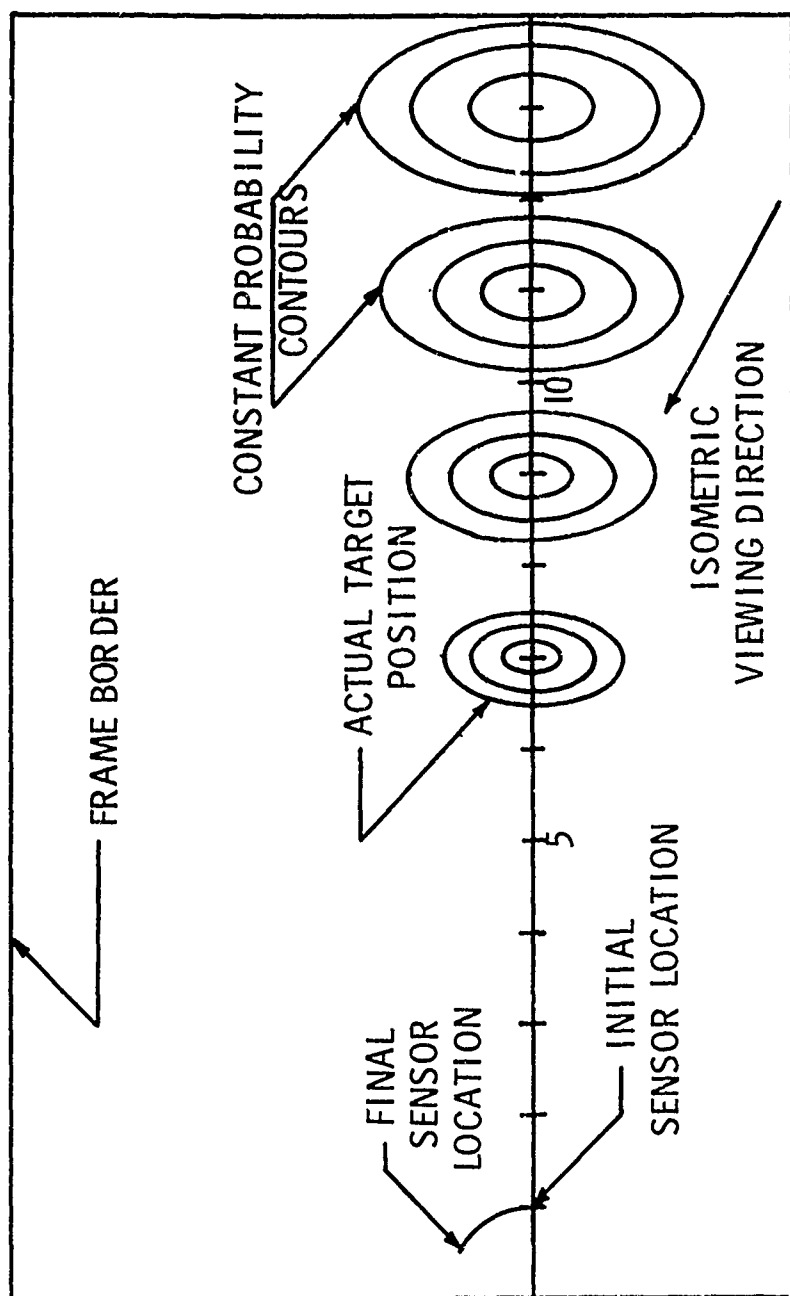
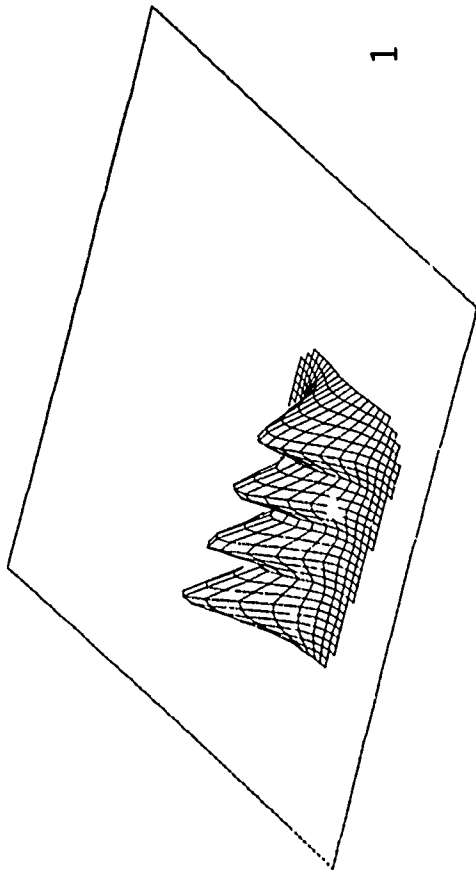


Fig. C-2. A Priori Density Resulting from  
Multipath Detection Ambiguities

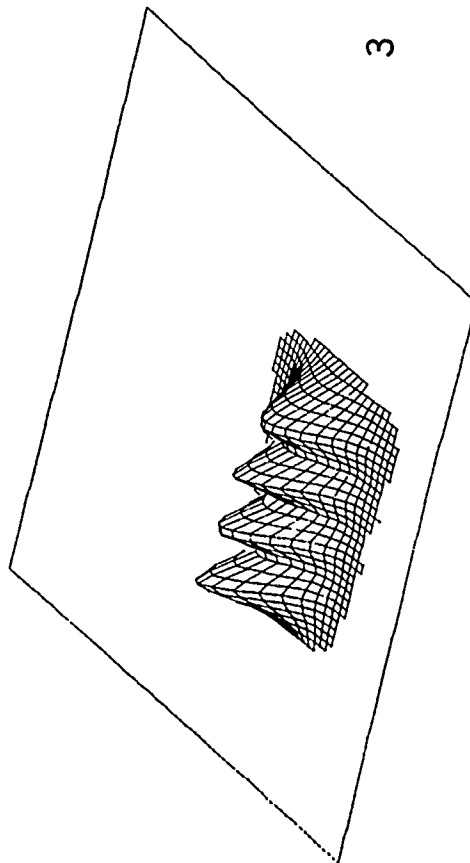
Fig. C-3. A Typical Sample Path  
Resulting from the Multipath Detection Ambiguity

The following 21 pages constitute Fig. C-3. Note the initial error mode lock on and then the gradual recovery to the correct mode.

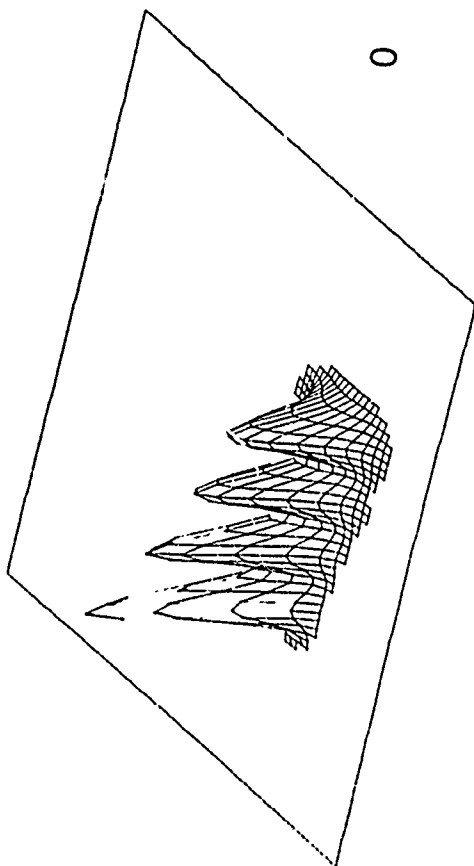
1



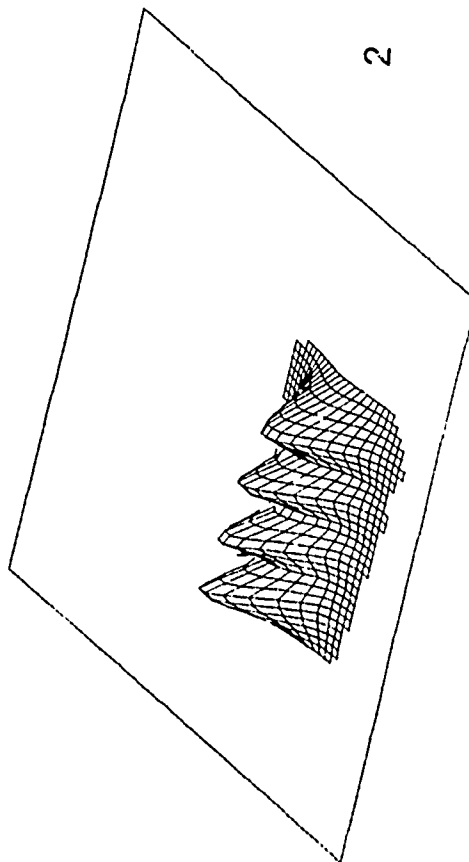
3



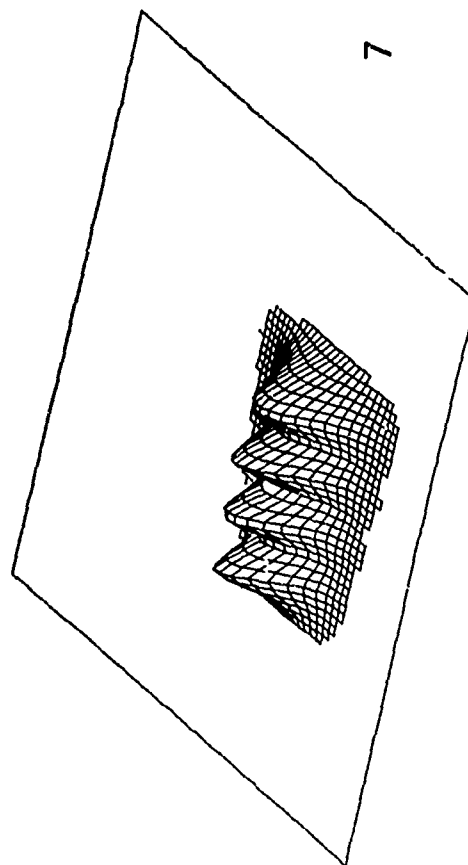
0



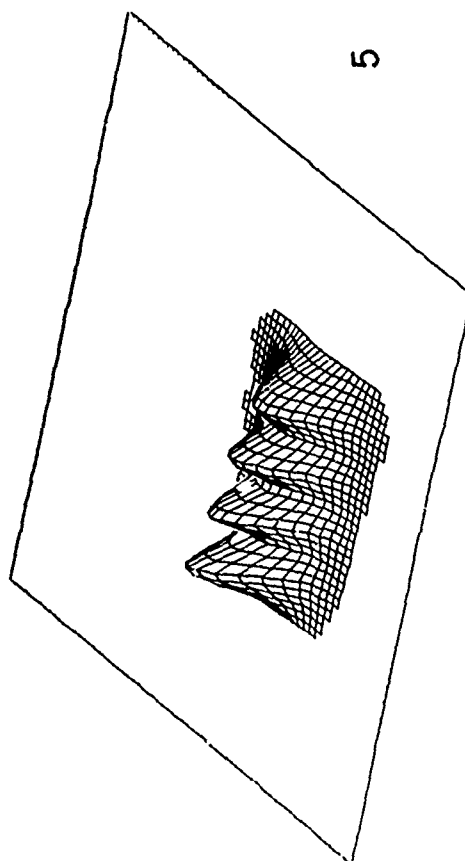
2



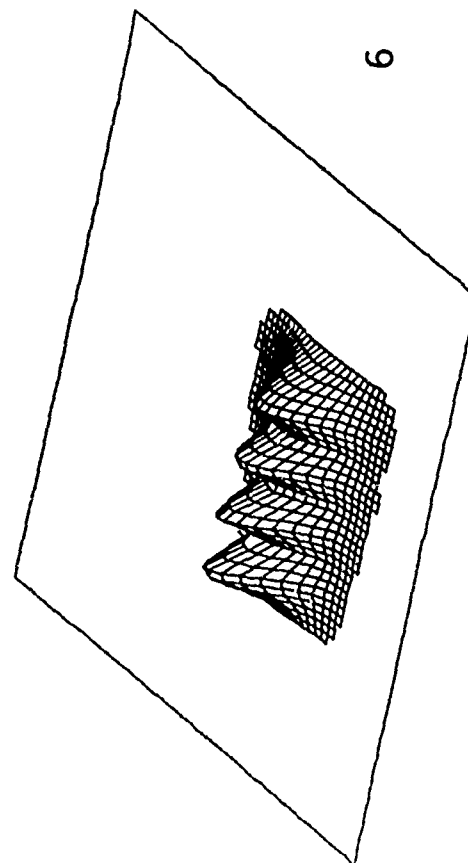
7



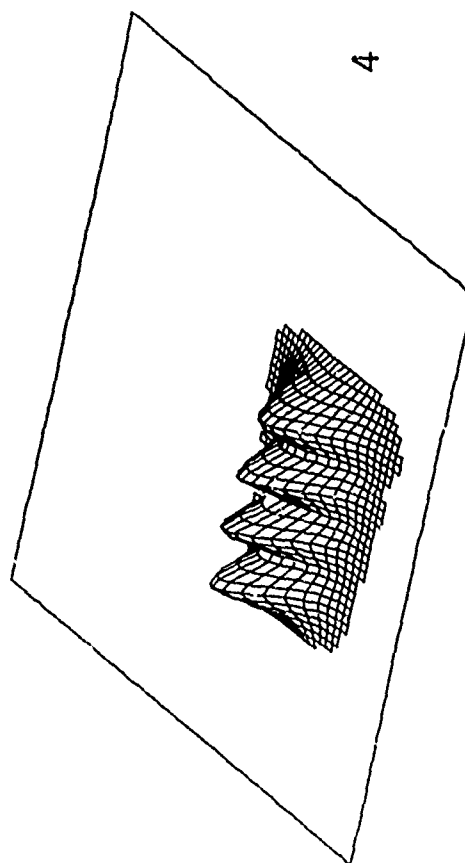
5

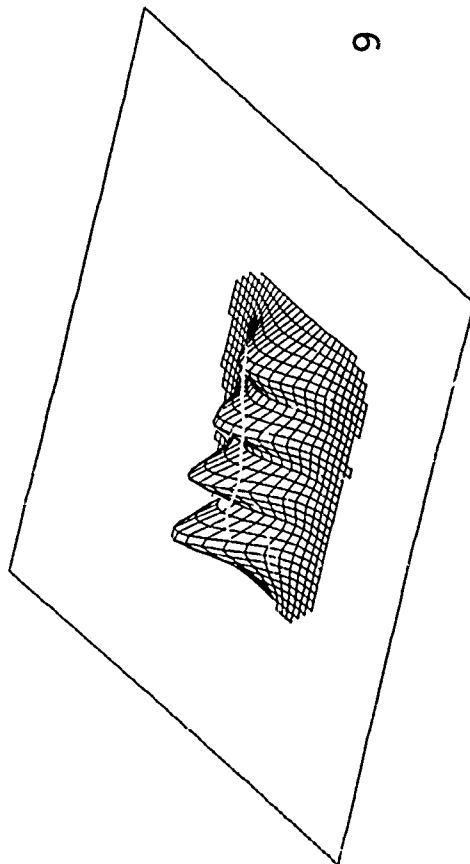


6

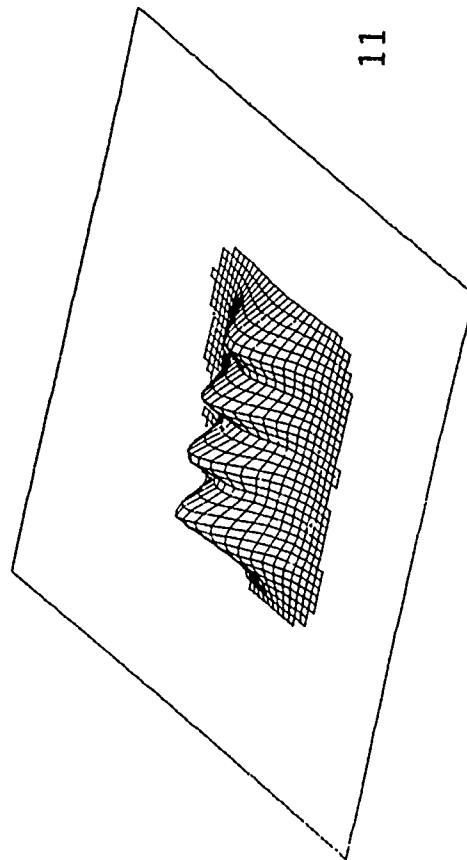


4

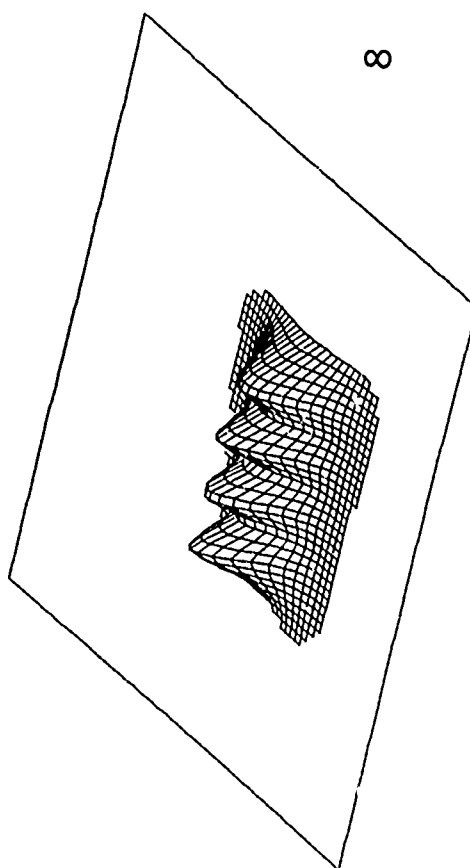




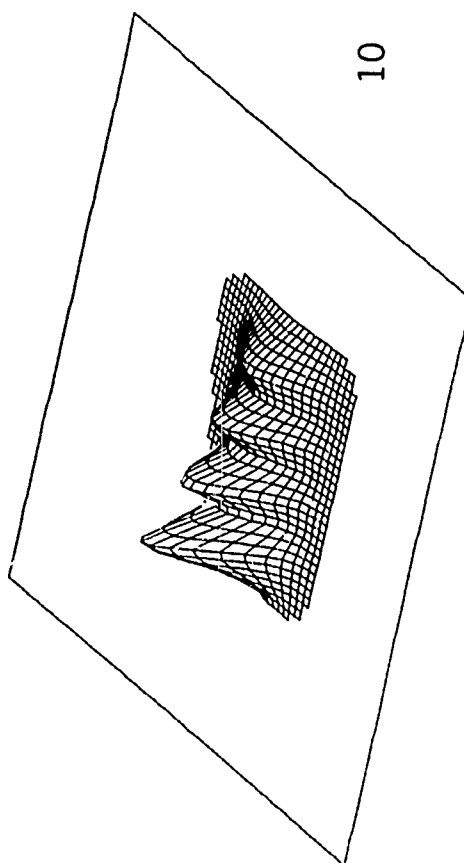
9



11

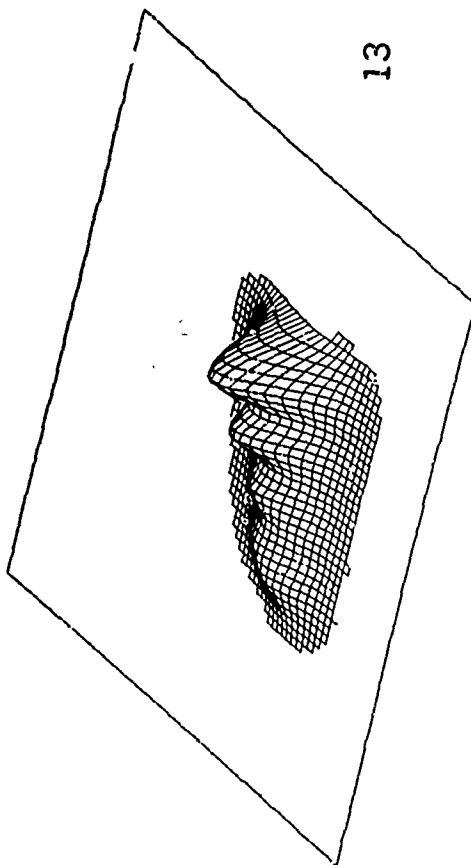


8

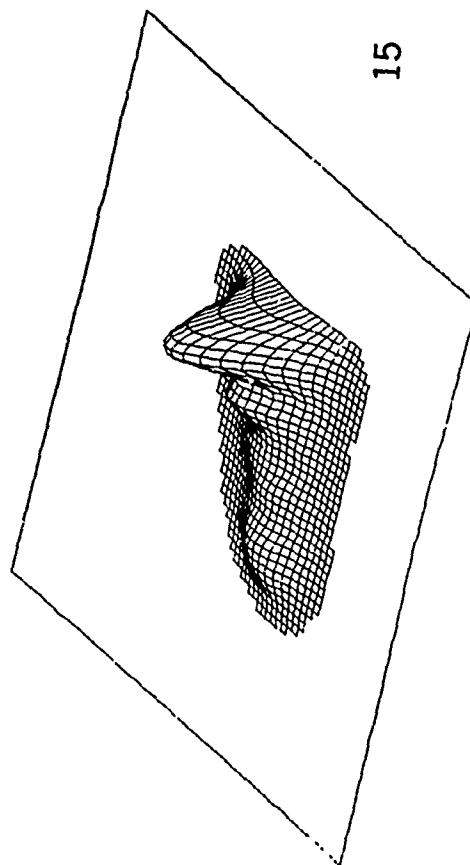


10

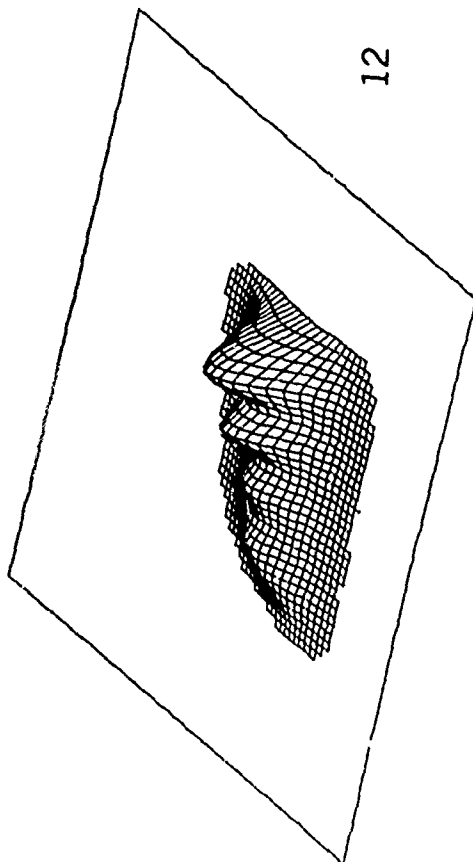
13



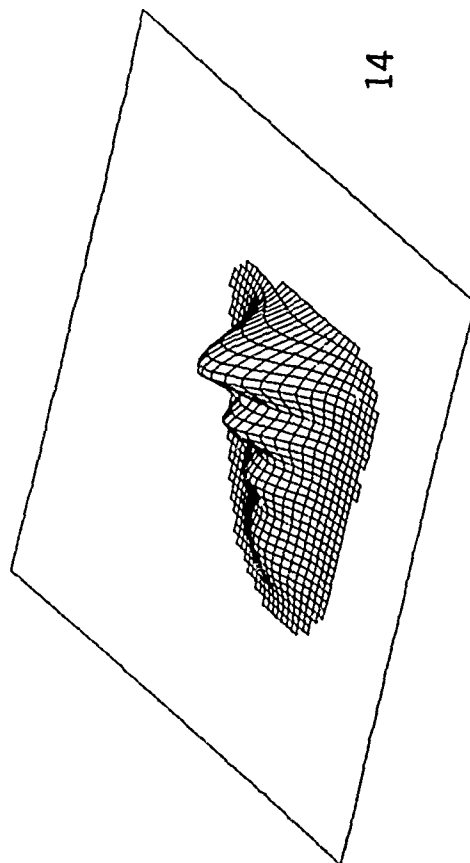
15

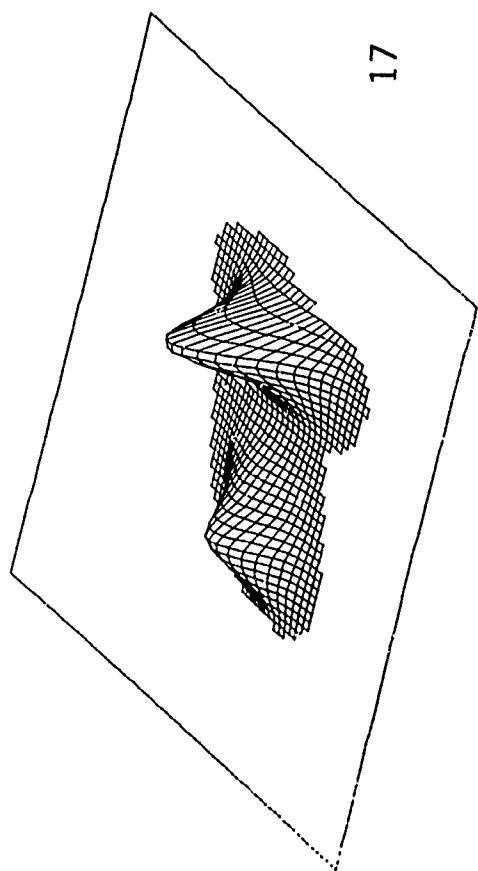


12

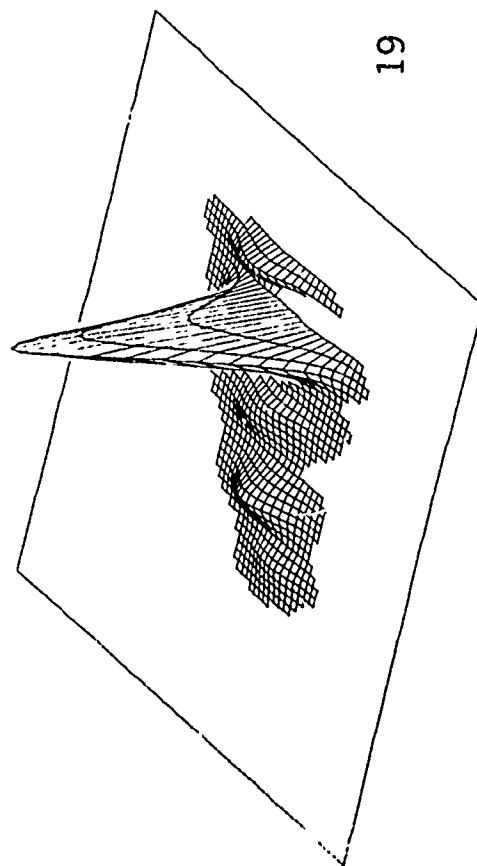


14

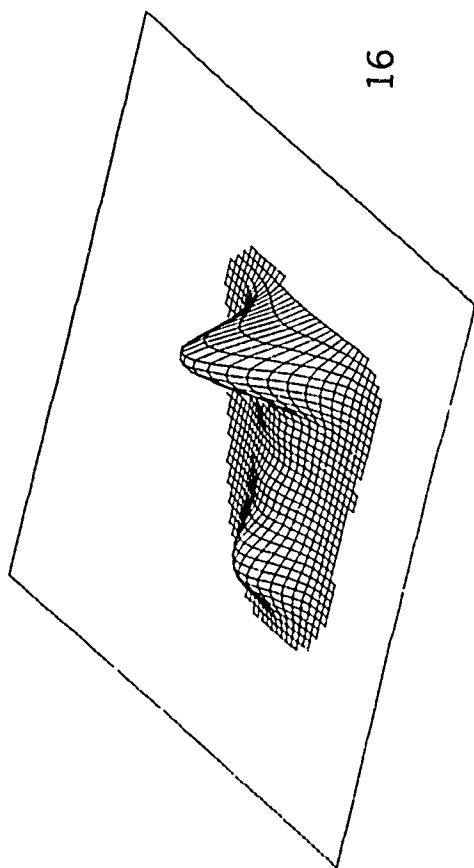




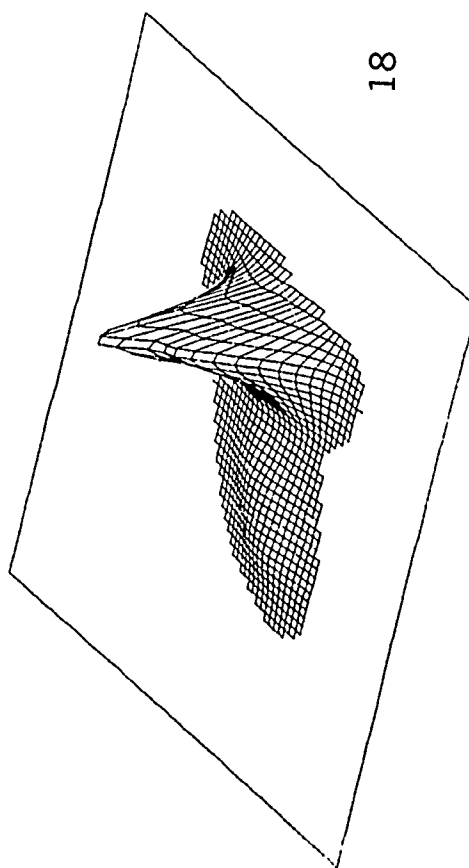
17



19

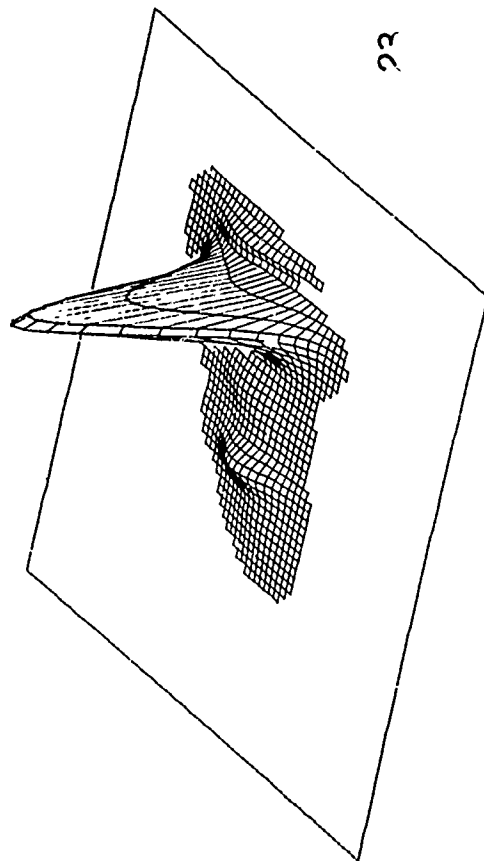


16

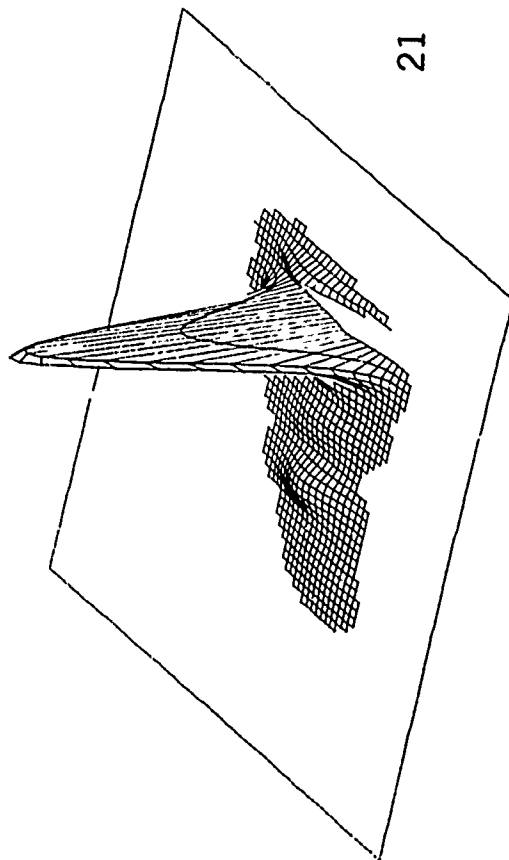


18

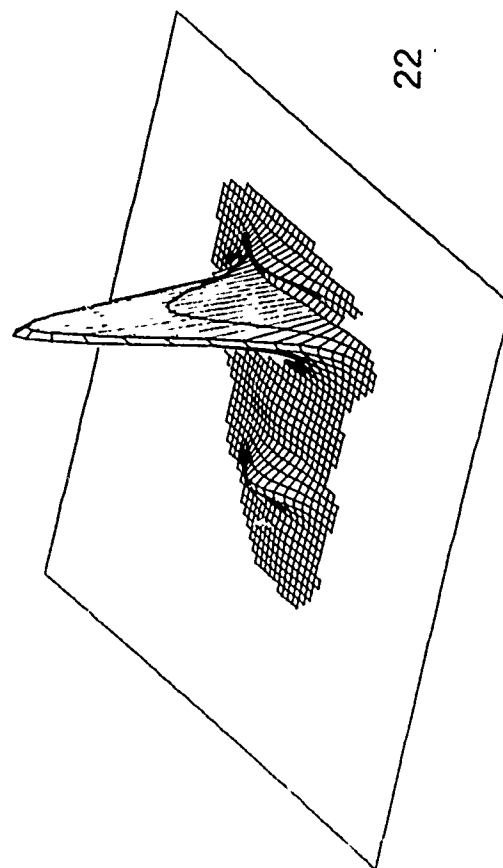
23



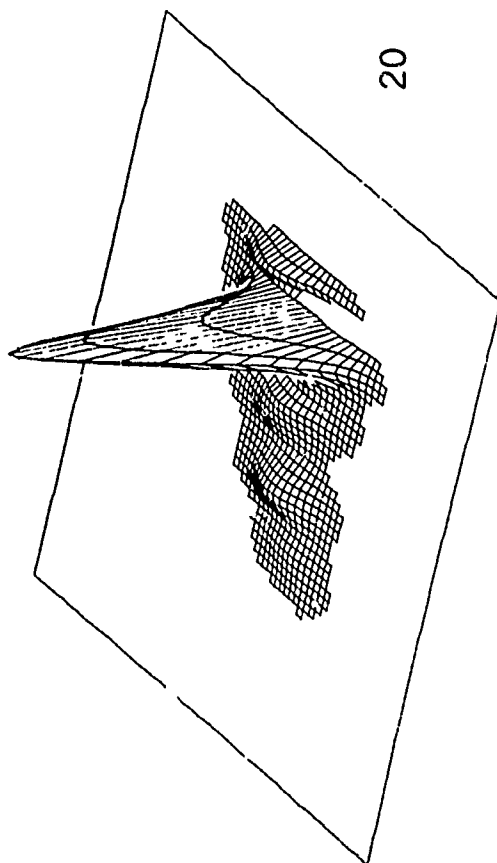
21

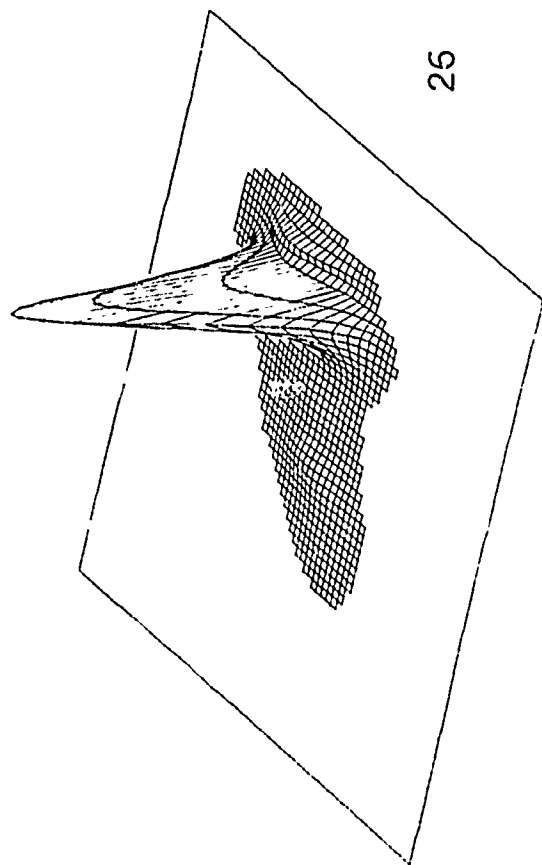
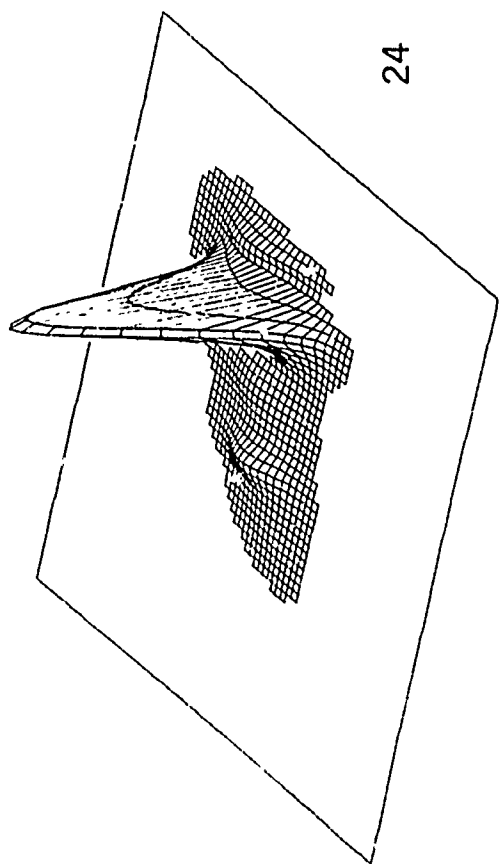
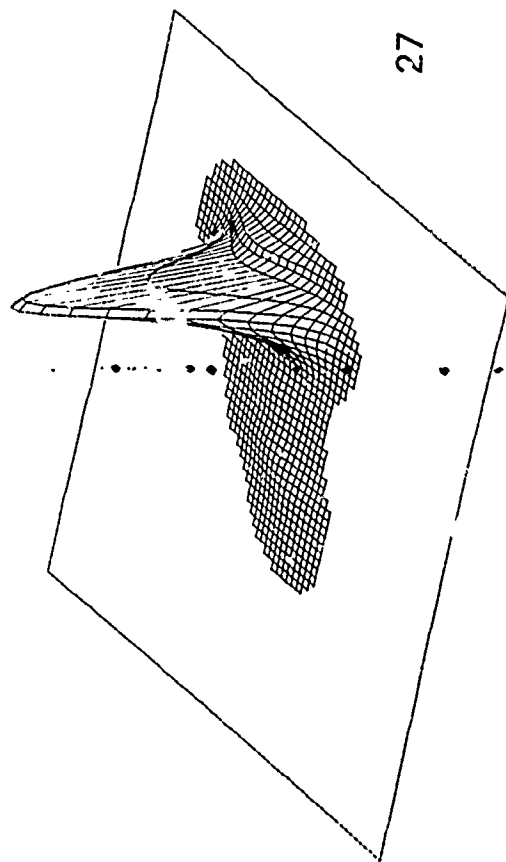
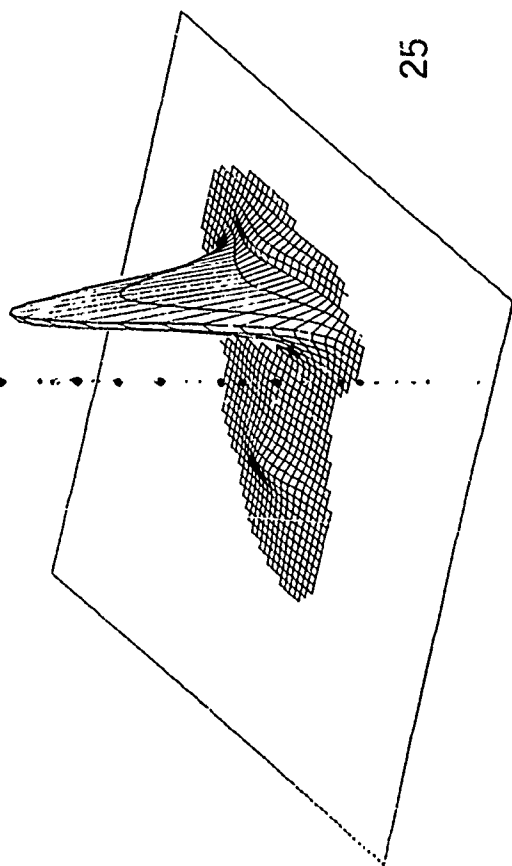


22

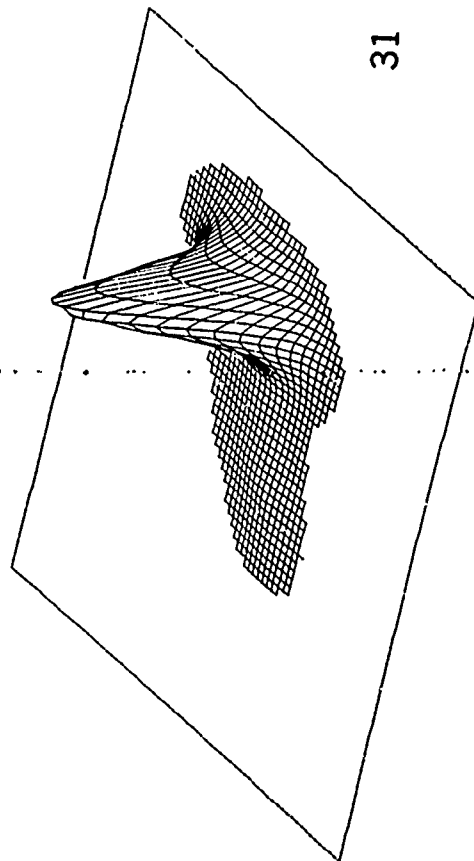


20

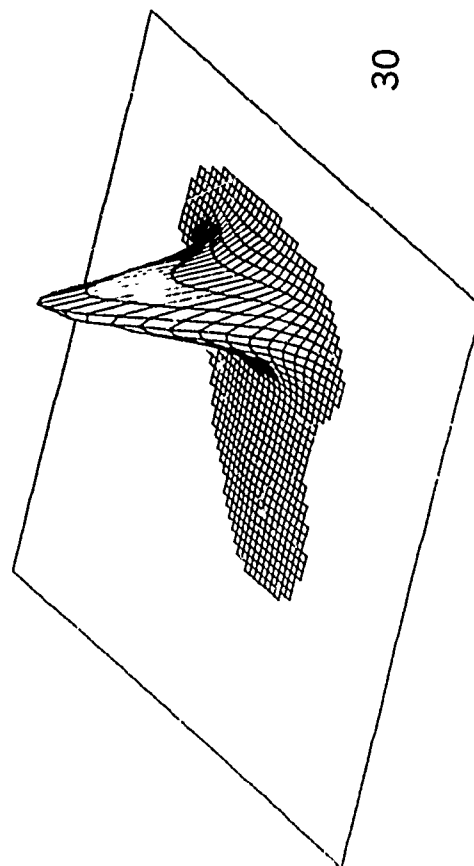




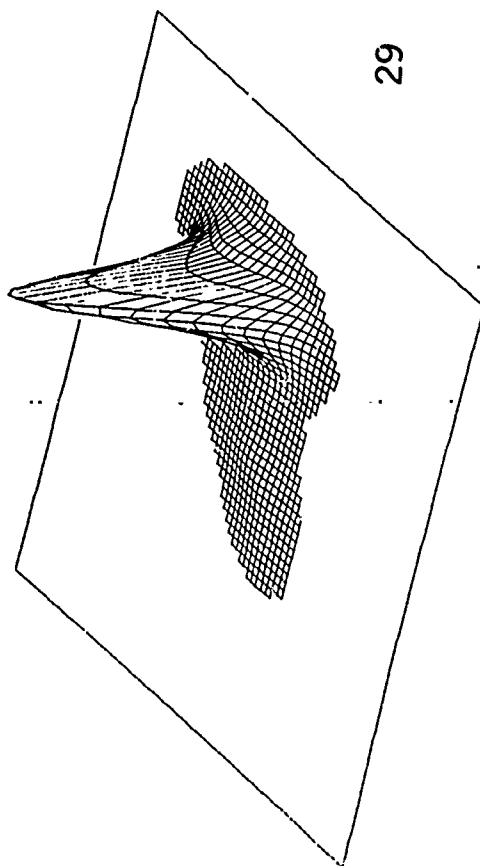
31



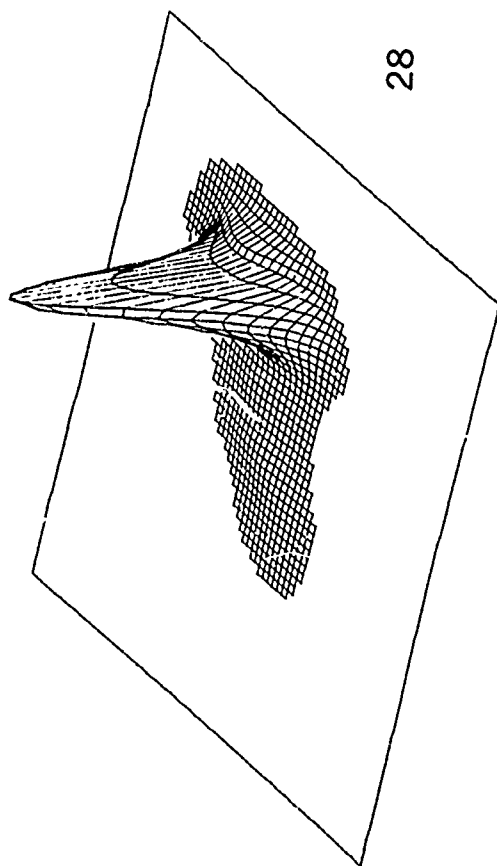
30

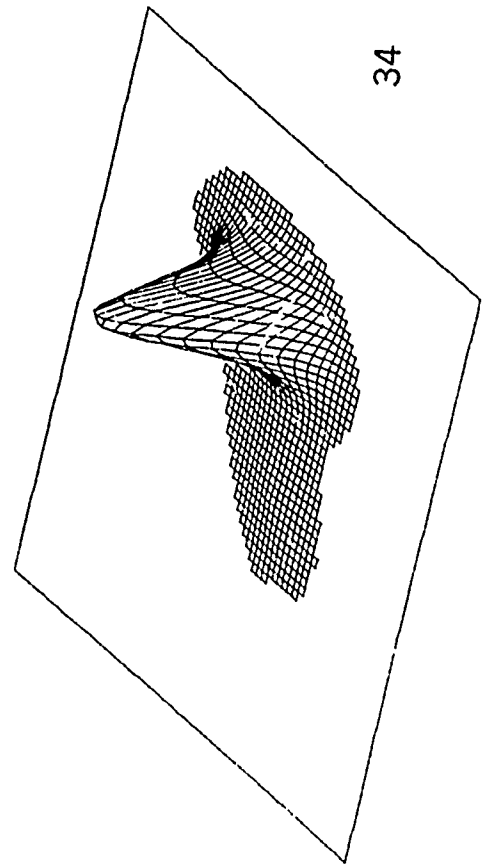
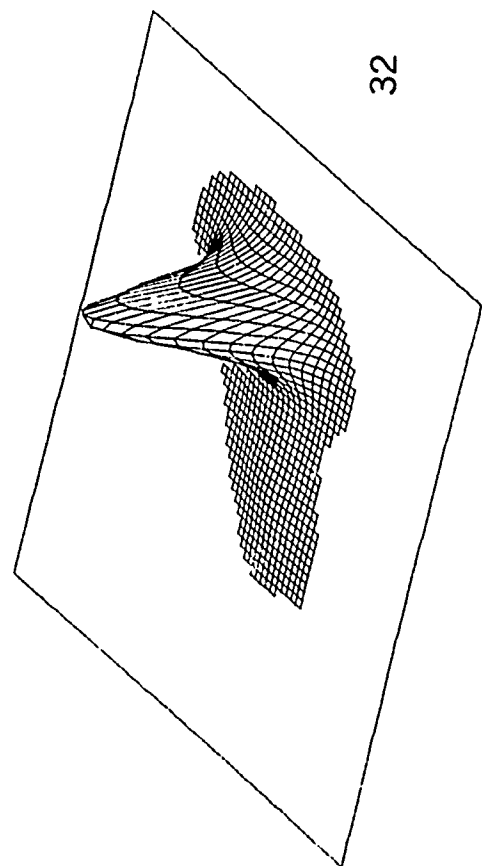
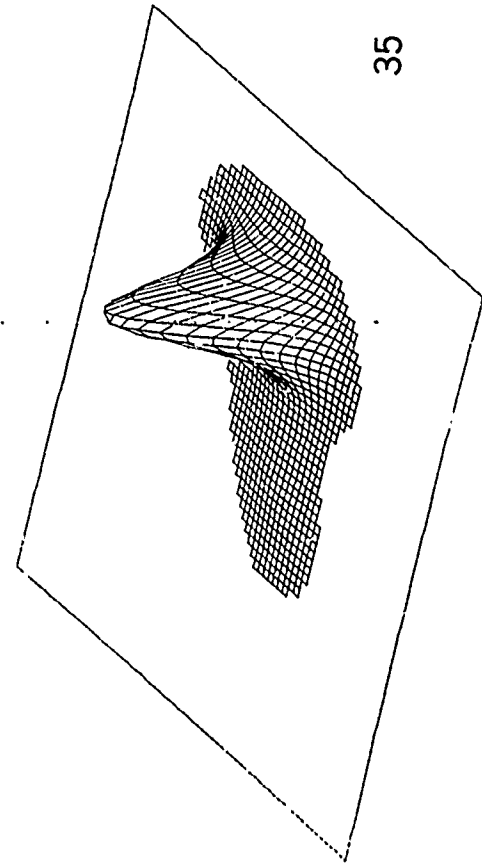
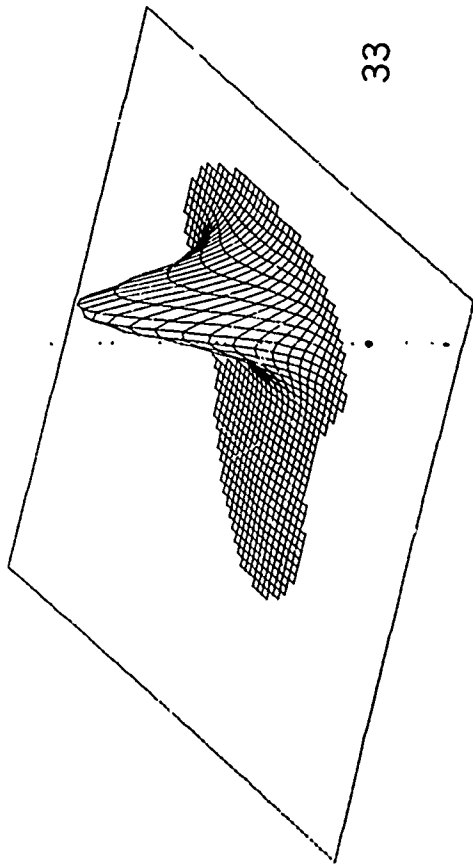


29

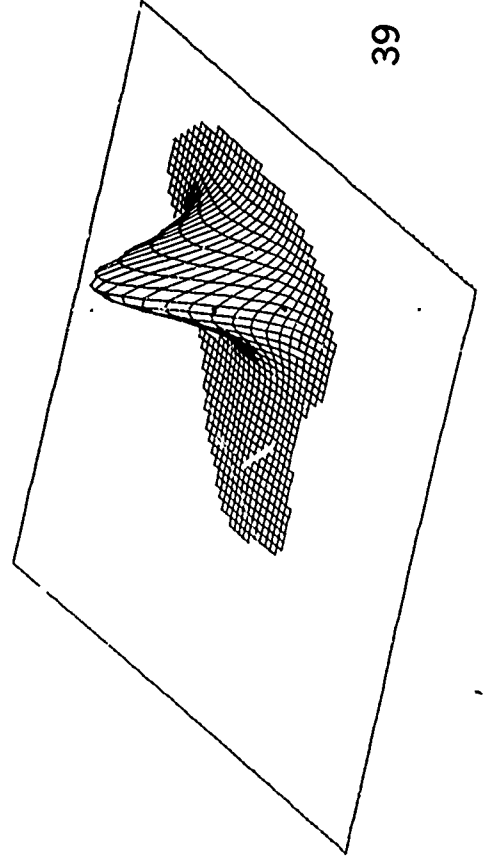


28

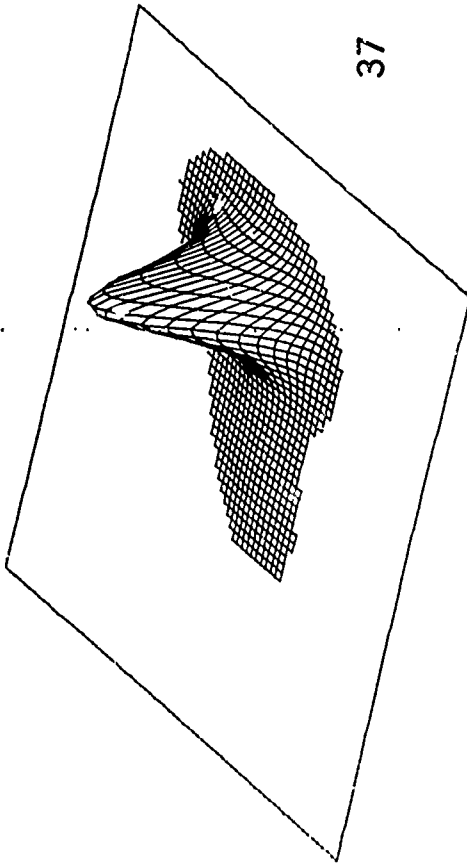




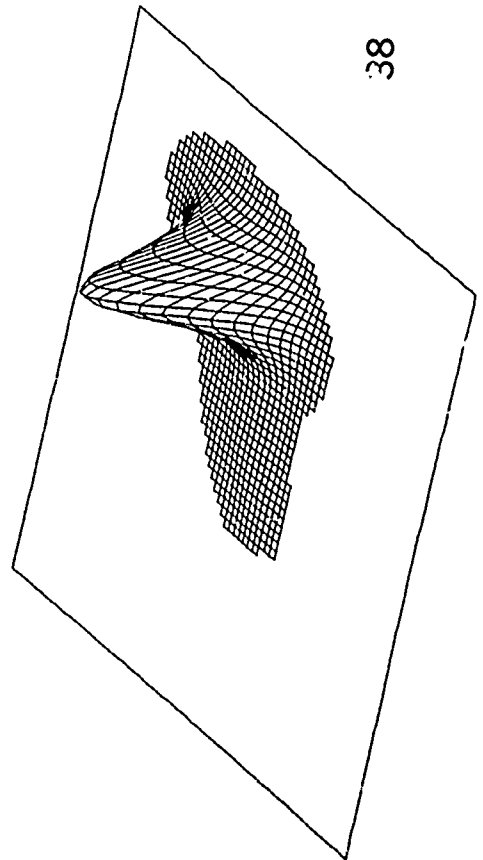
39



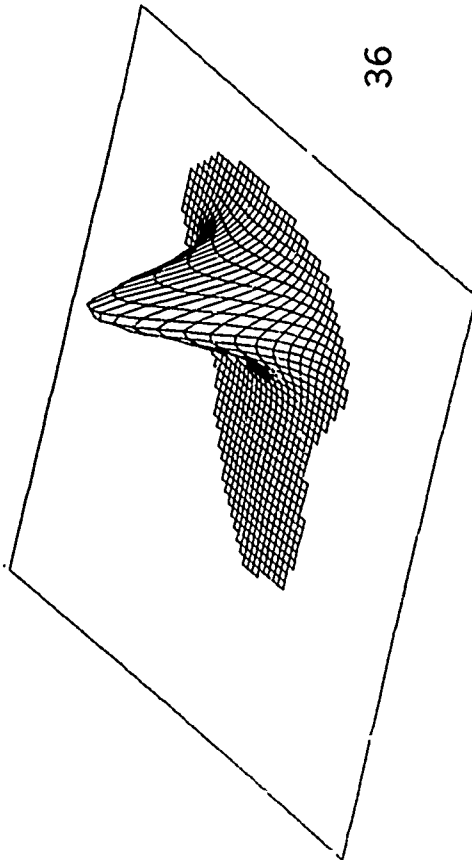
37

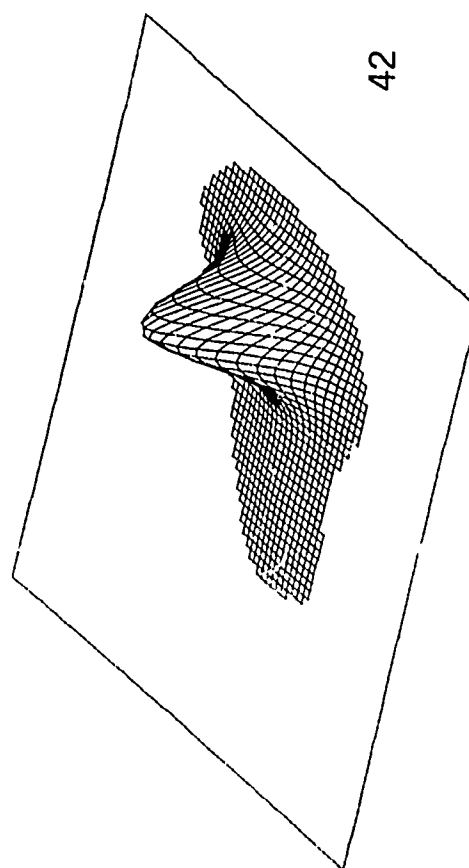
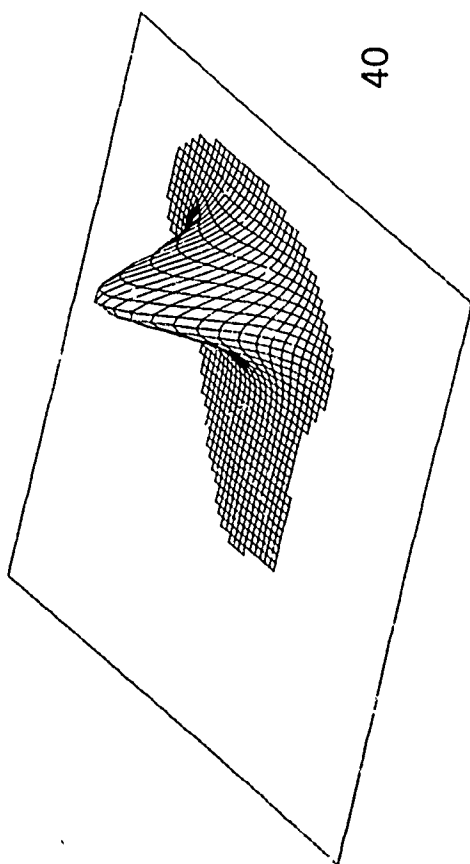
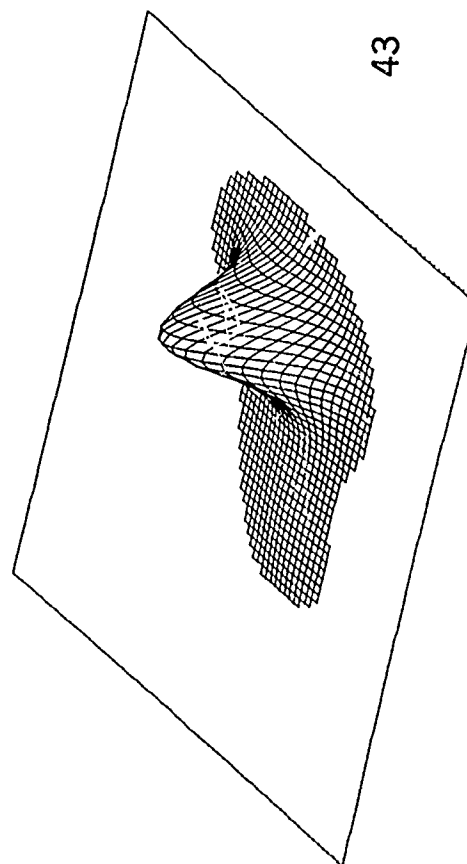
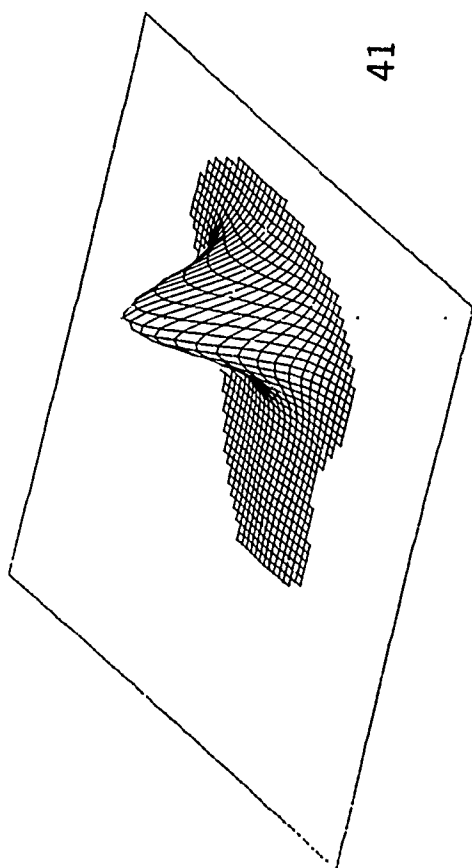


38

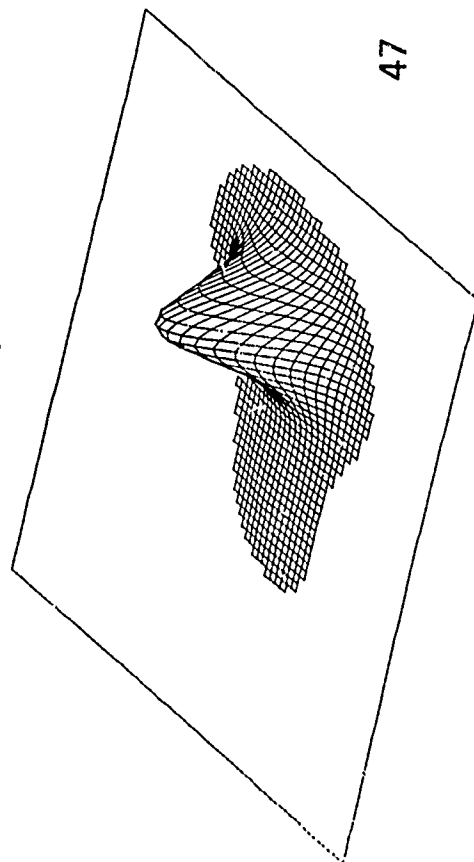


36

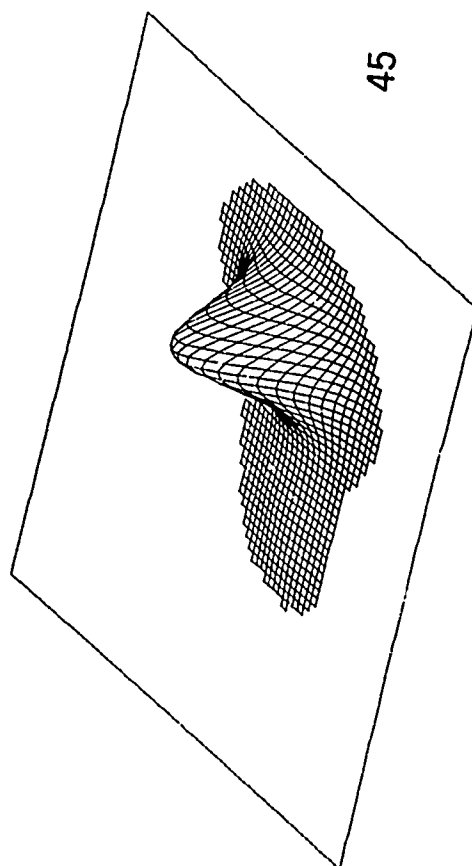




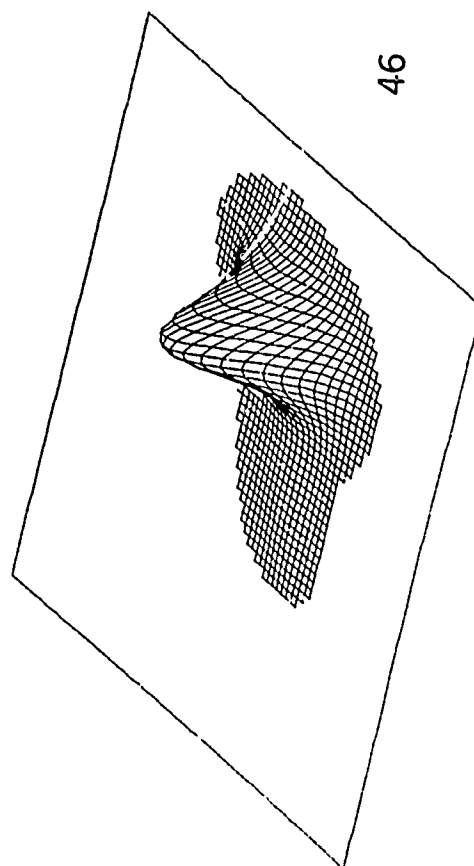
47



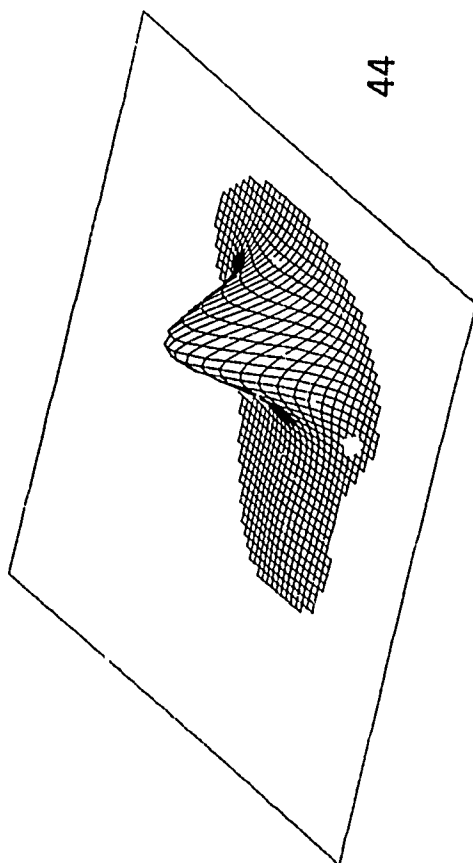
45

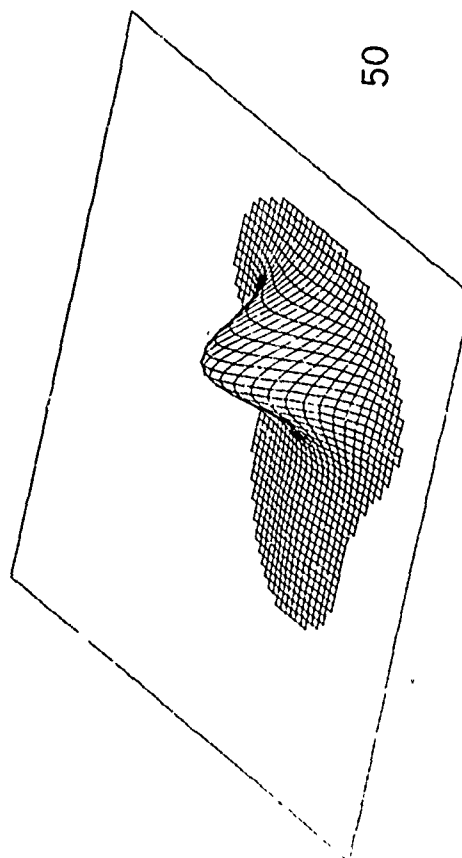
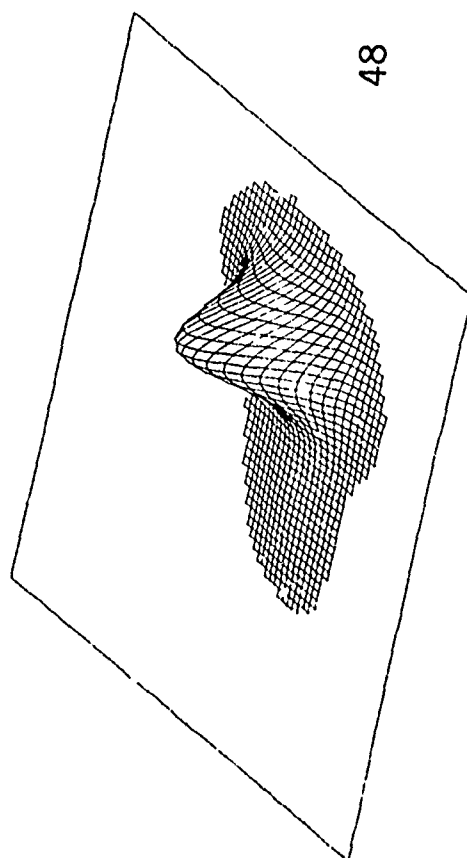
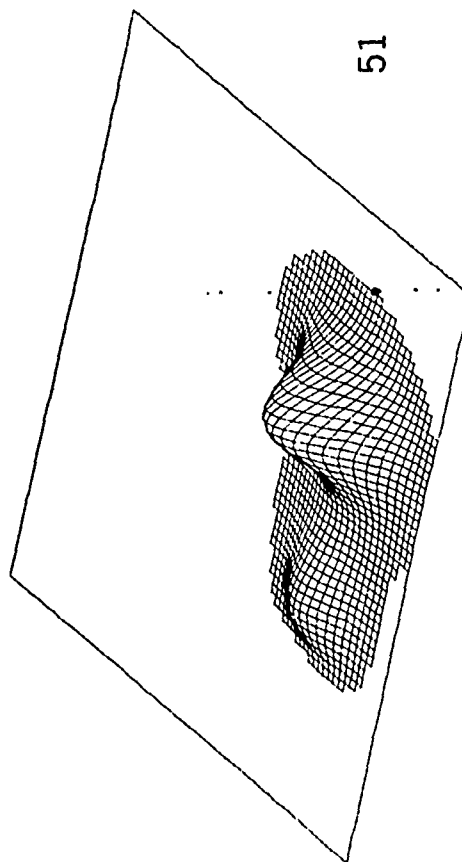
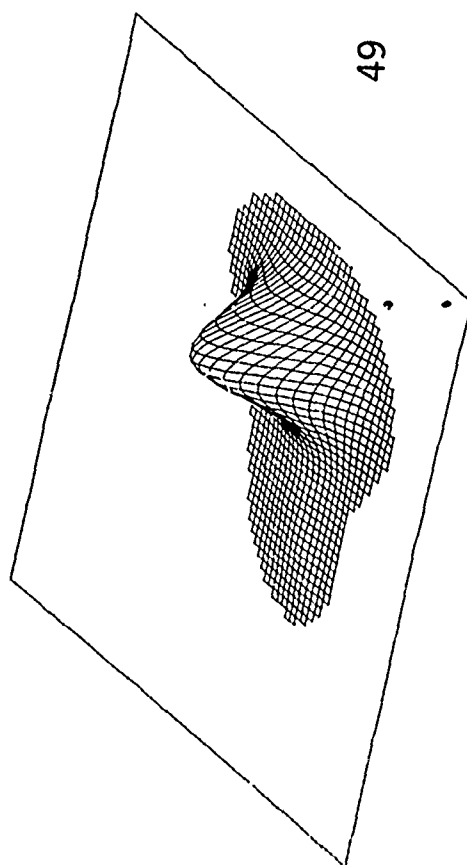


46

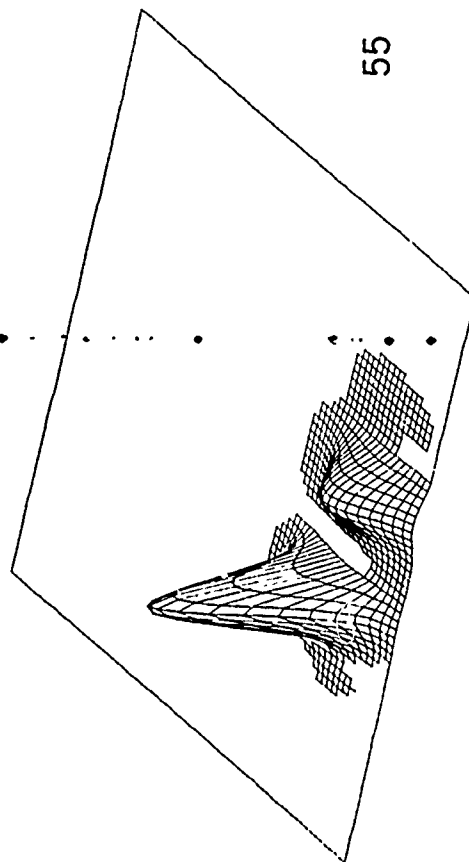


44

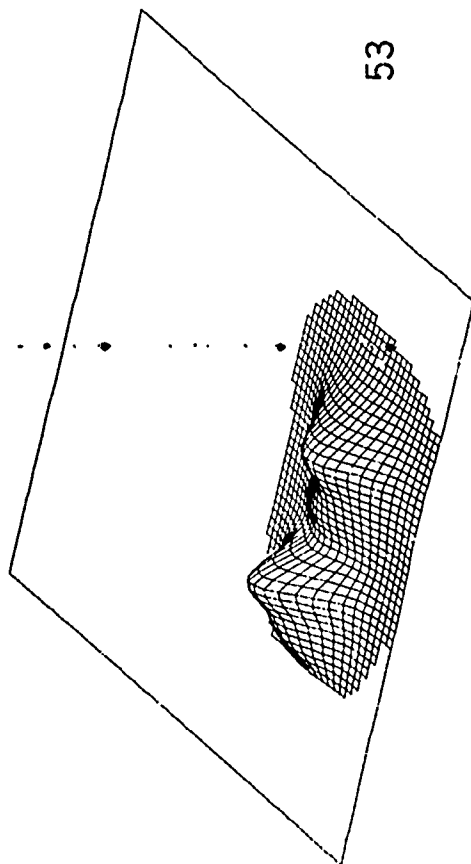




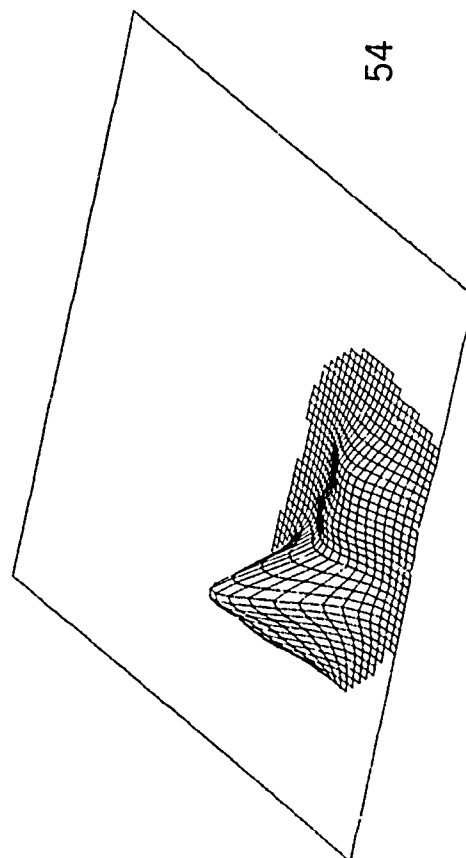
55



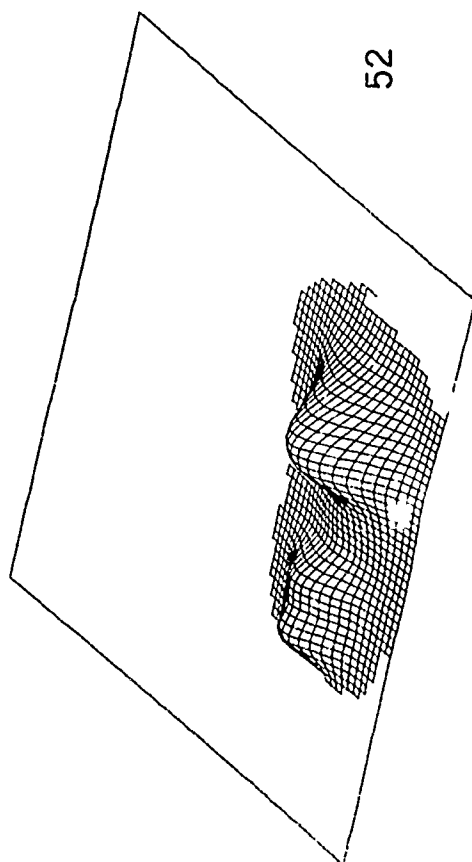
53

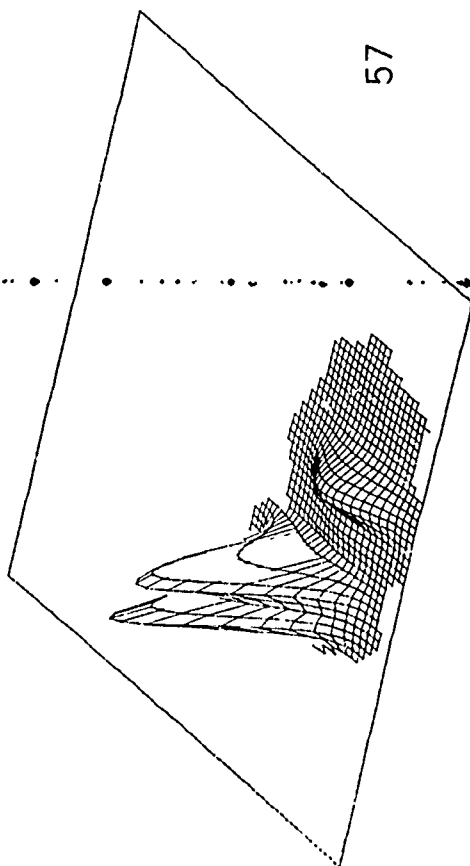


54

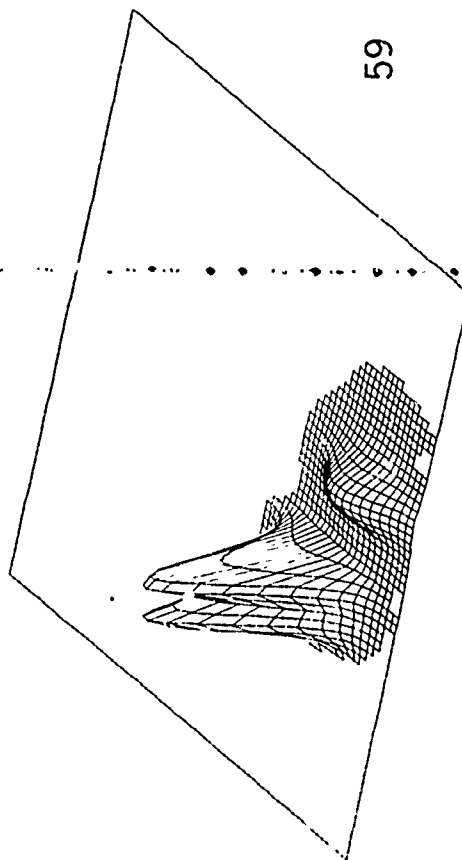


52

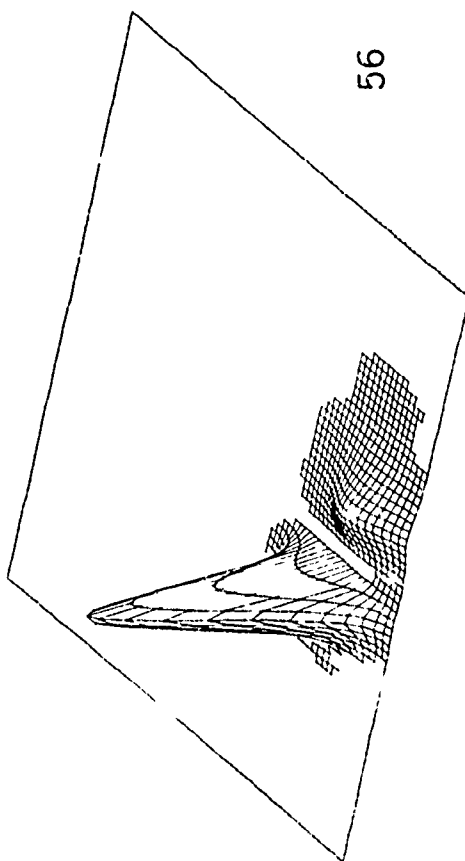




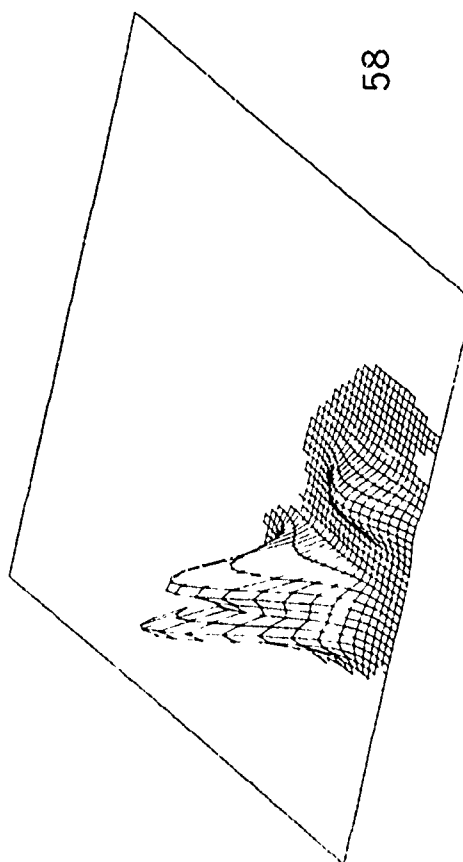
57



59

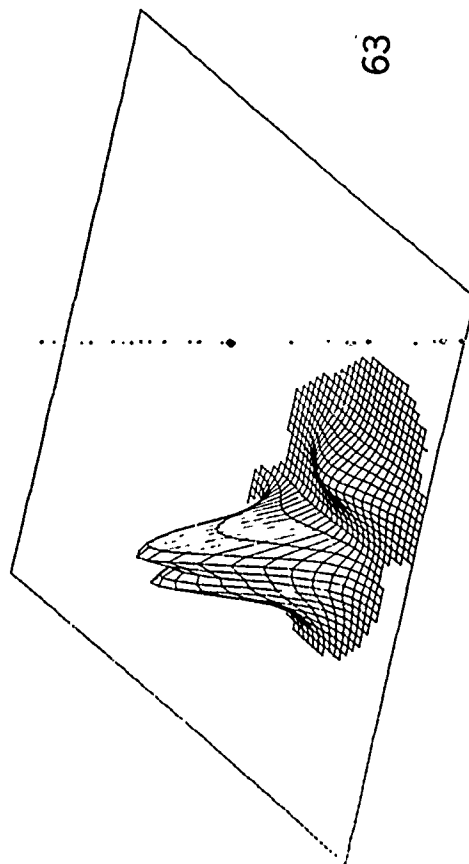


56

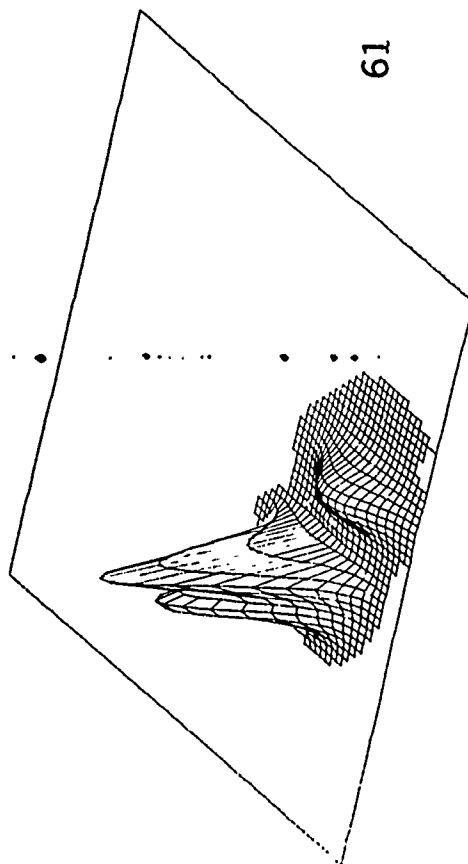


58

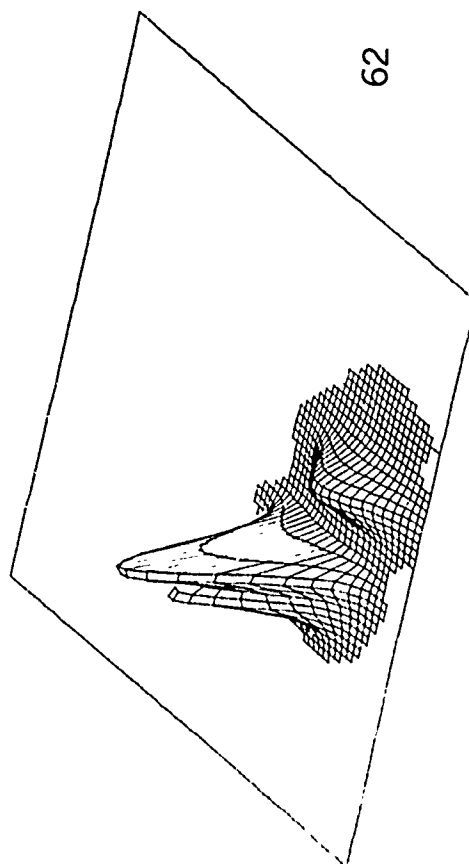
63



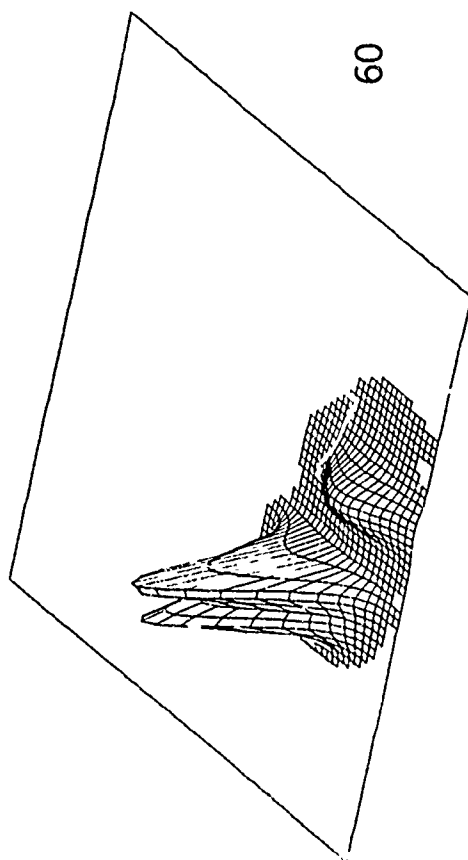
61

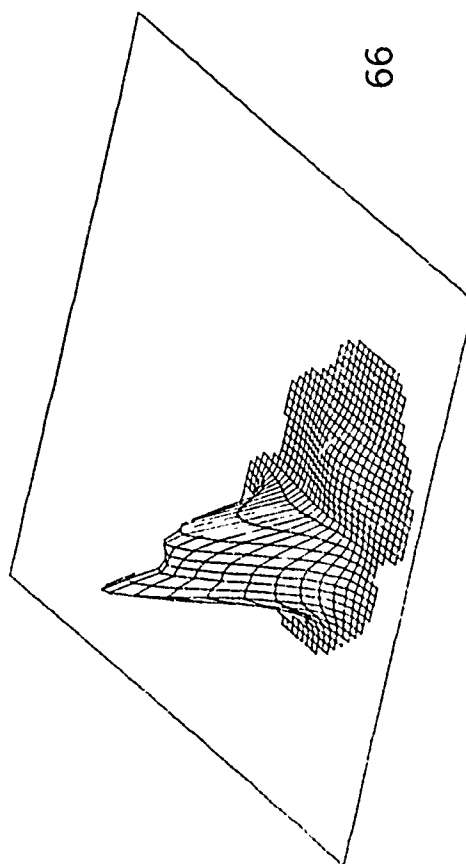
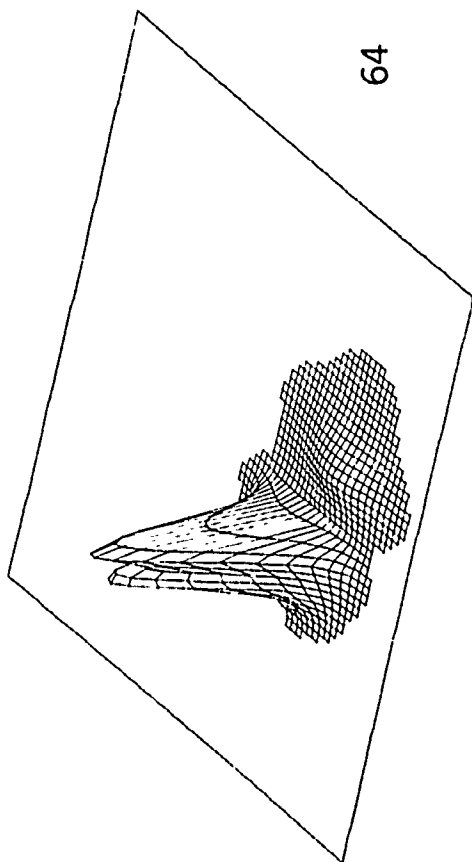
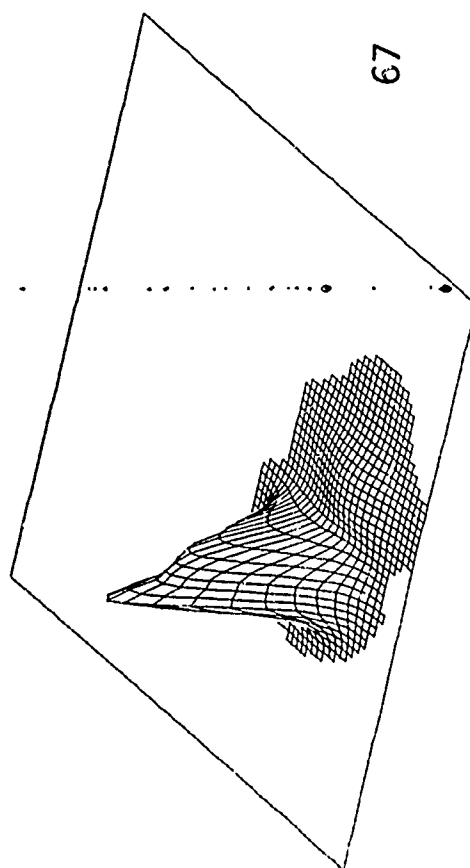
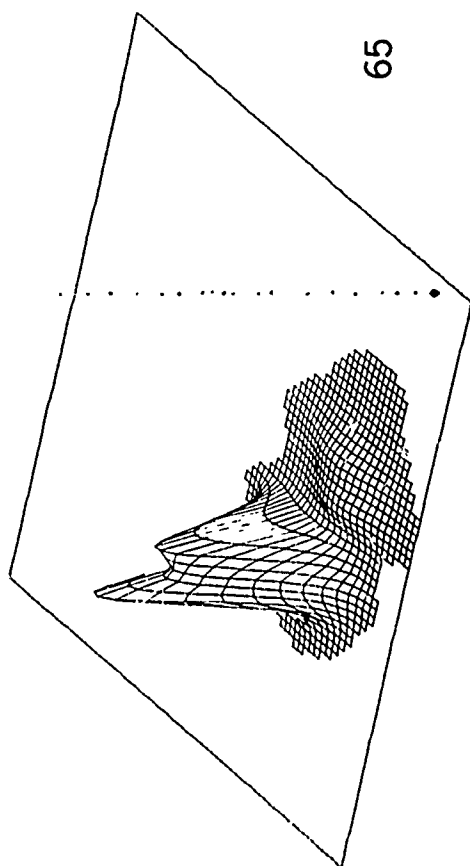


62

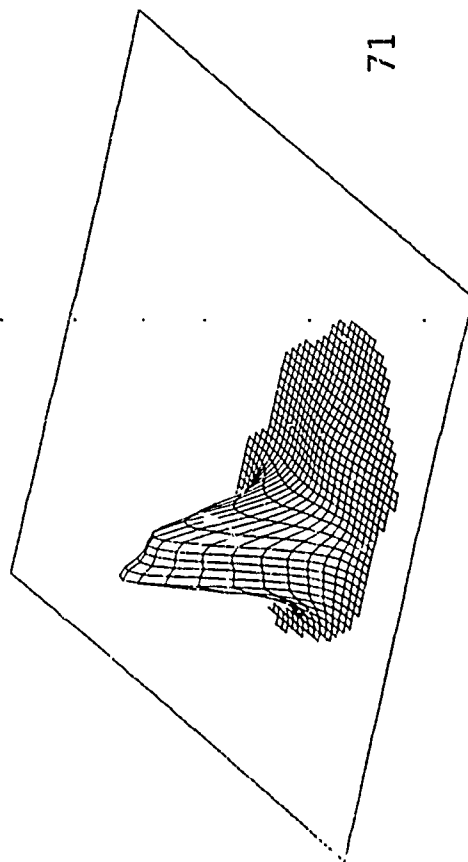


60

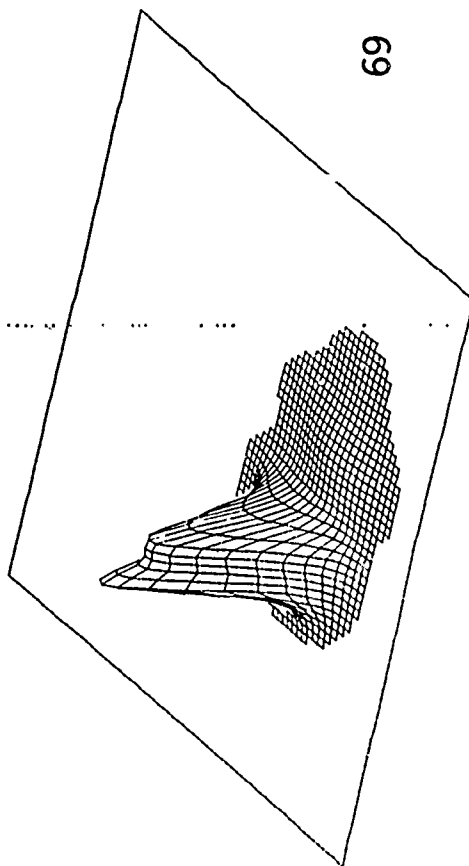




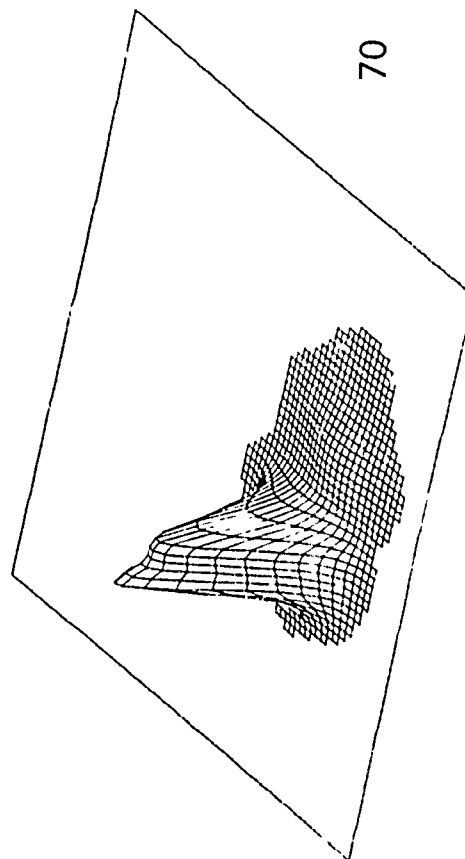
71



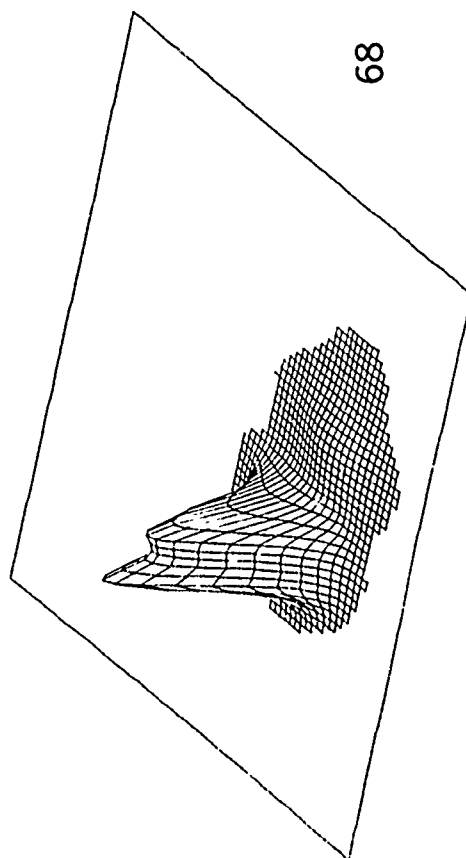
69

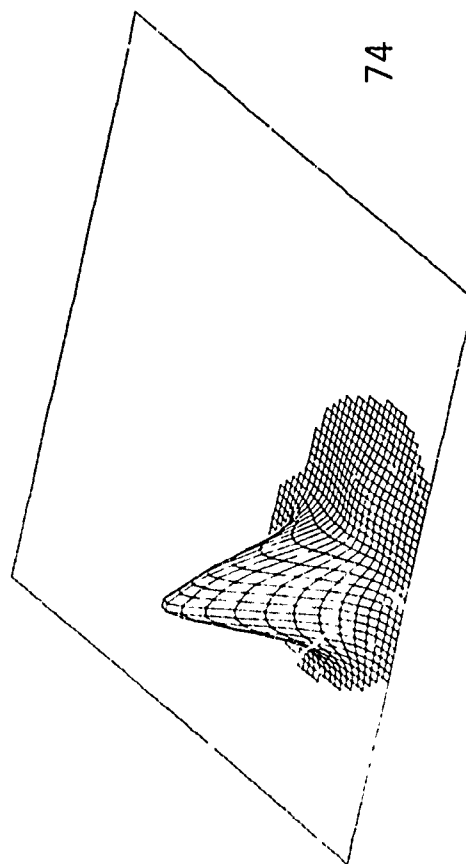
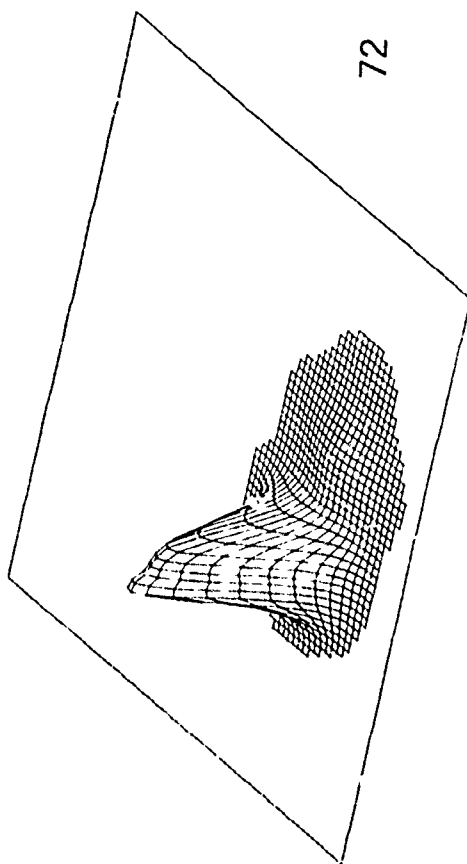
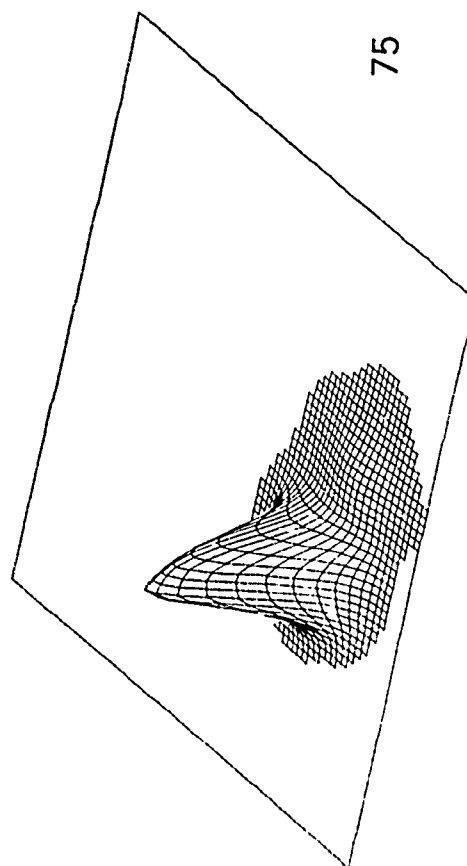
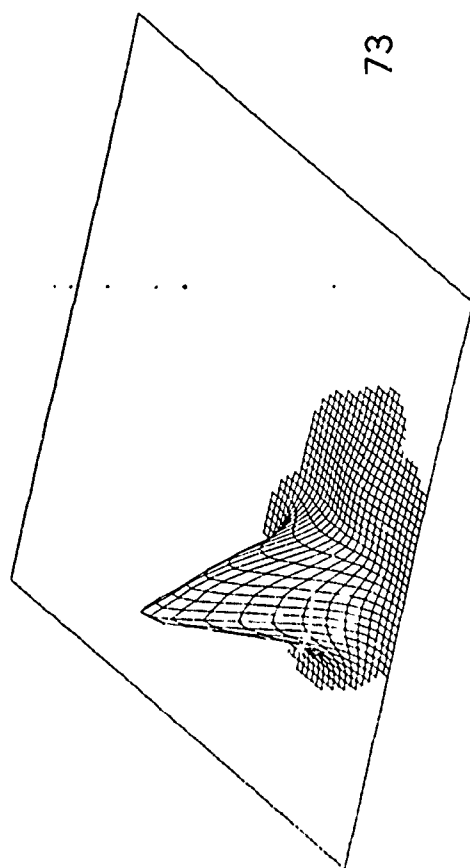


70

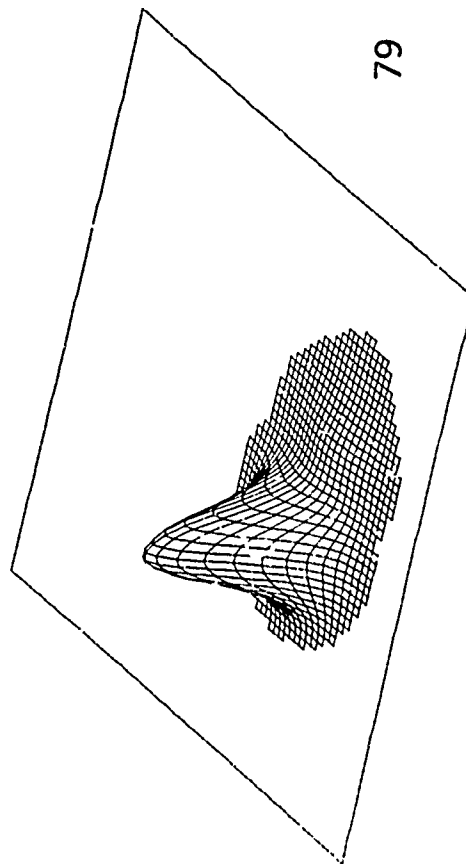


68

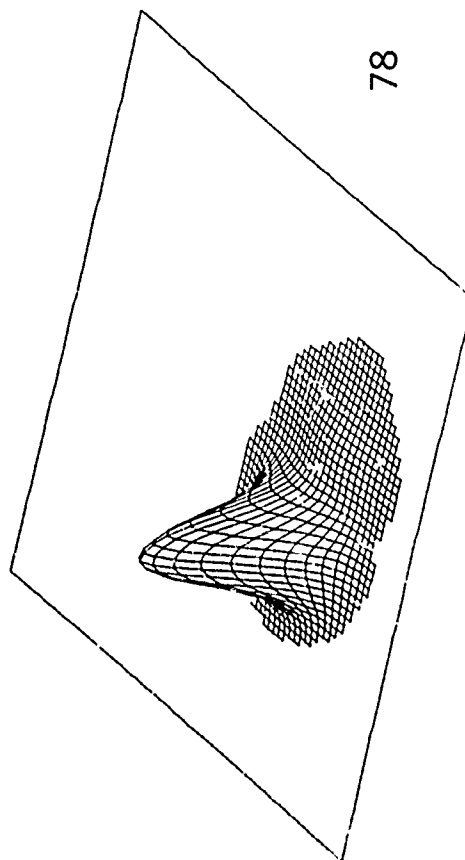




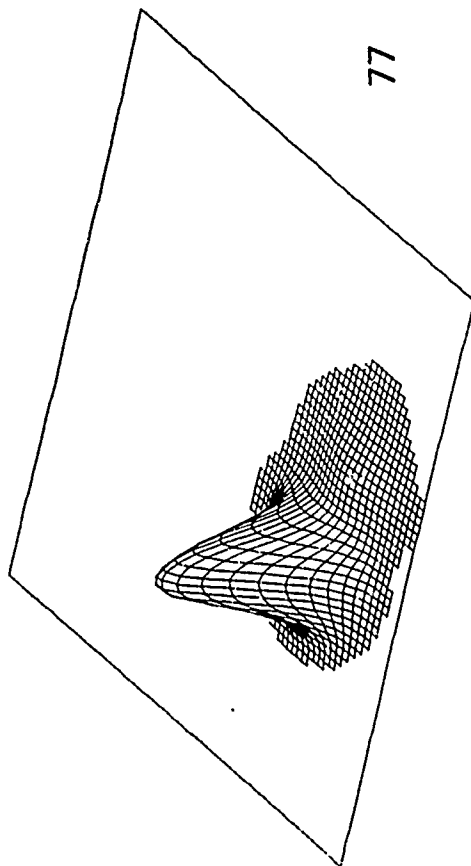
79



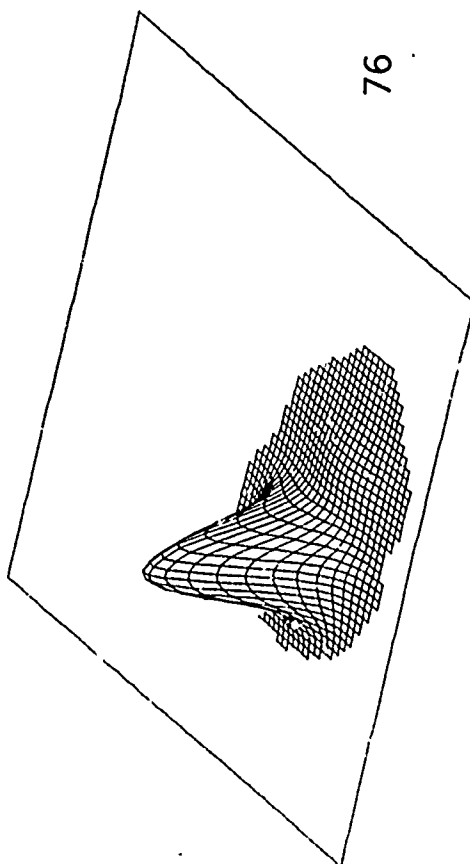
78

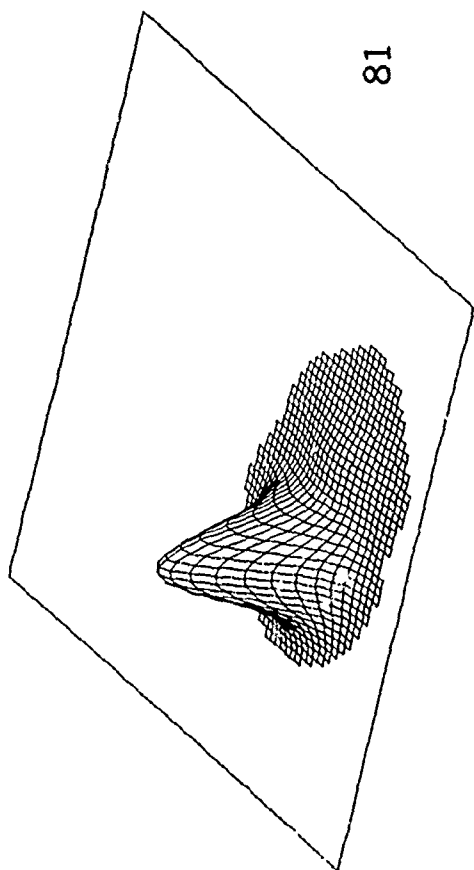


77

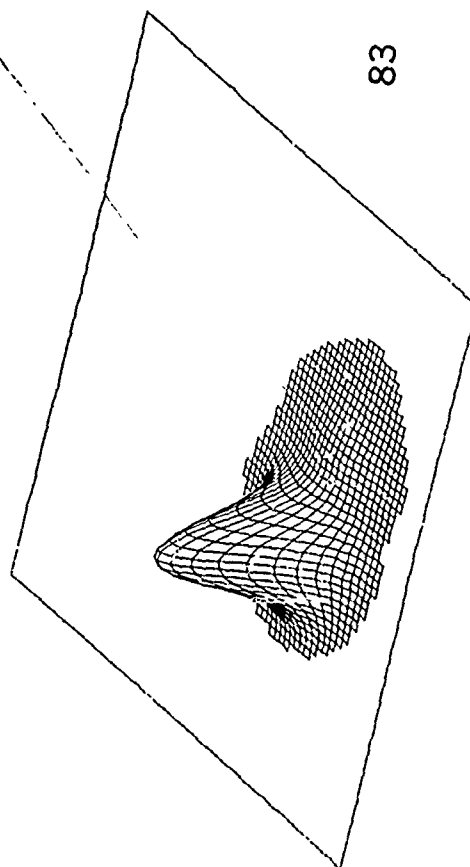


76

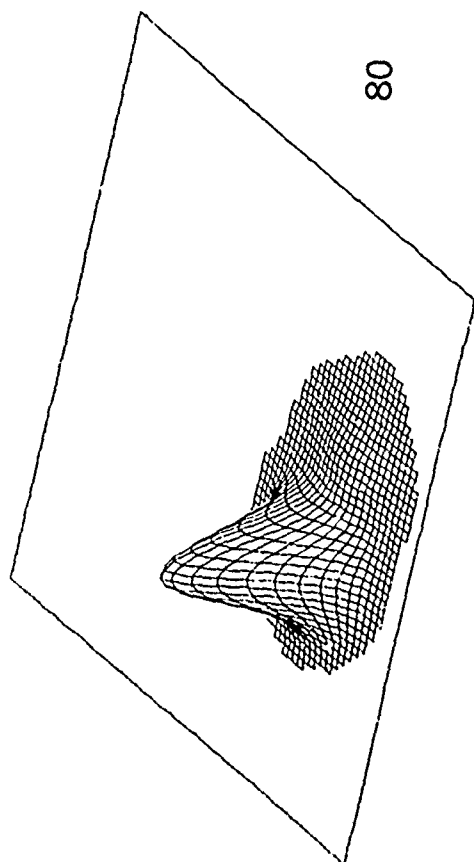




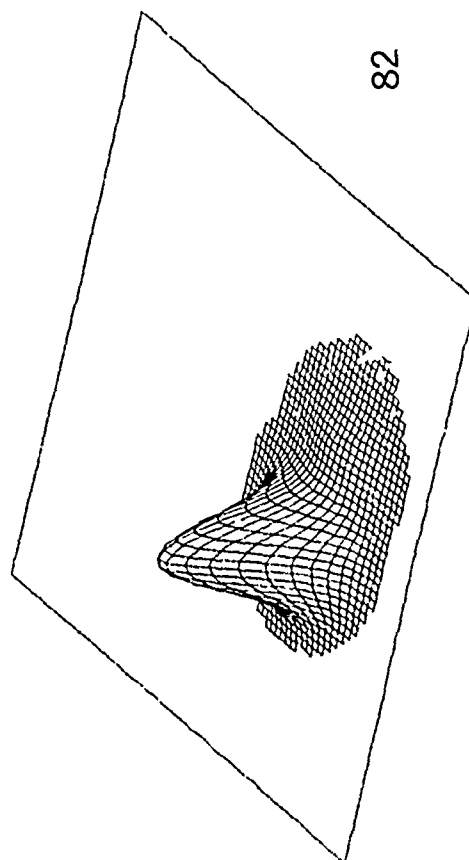
81



83



80



82

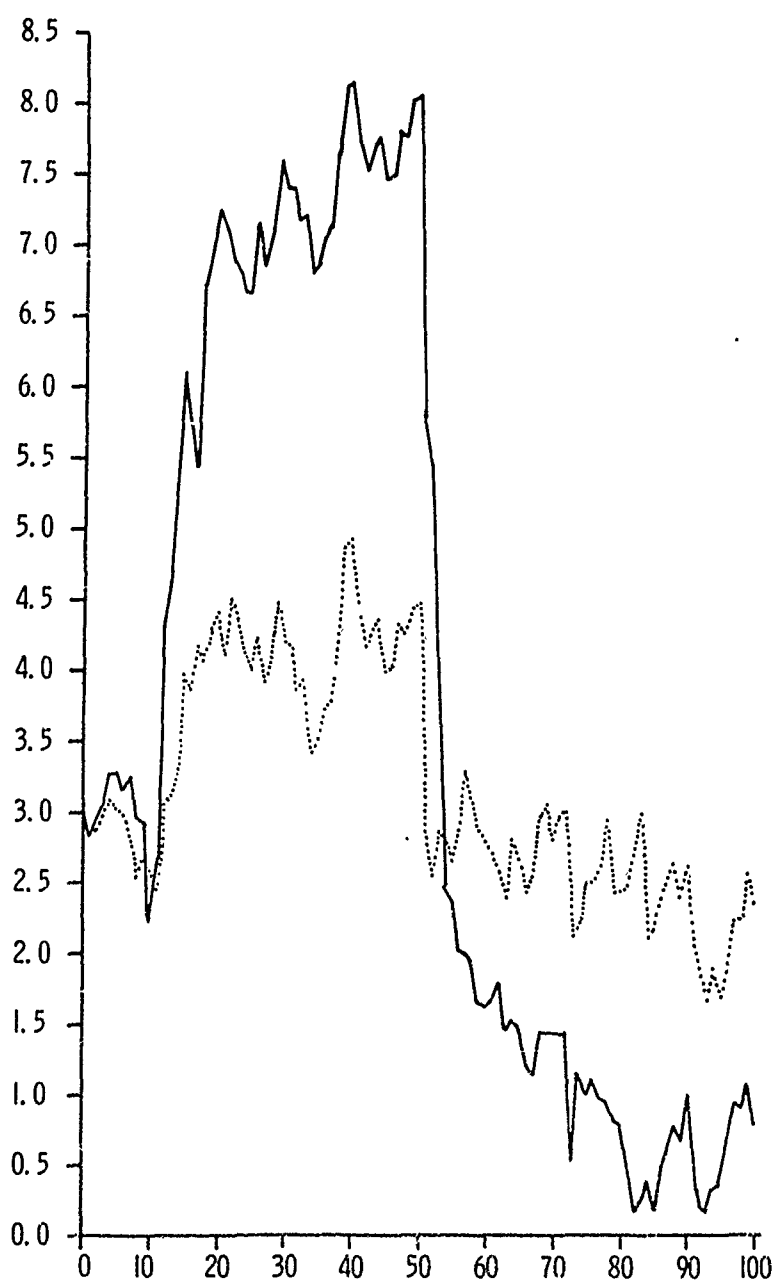


Fig. C-4. Absolute Error Performance of Optimal and Linearized Predictors for Multimodal Problem

The examples have shown the wealth of experimental experience available to the estimation engineer as a result of numerical studies of the problem. It is possible, for example, to assess the validity of given analytical approximations and the value of unused information regarding the structure of the estimation problem. In addition it is possible to compute practical lower bounds on performance for nonlinear estimators in some small dimensional problems, a fact which may have some important engineering significance to designers of sensors, for example, who must know the theoretical limitations on the performance of a given design. It is anticipated that future experiments along these lines will substantiate this conjecture. In the meantime there appears to be many possible application areas to explore. See, for example, Chapter VII.

## VII. Example: Optimal Nonlinear Phase Demodulation

### A. Introduction

An interesting application for optimal nonlinear estimation was introduced by Mallinckrodt, Bucy, and Cheng[7], who considered the problem of tracking a first-order phase process based on measurements of a modulated signal in noise of the form

$$dz(t) = A \cos [\omega_0 t + x_1(t)]dt + dv(t)$$

where  $A$  is a known amplitude,  $\omega_0$  is a known carrier frequency, and  $x_1(t)$  is the message process being tracked. The measurement noise is assumed white. Using a voltage-controlled oscillator the known carrier may be removed by heterodyning down to base band, producing both in-line and quadrature components and resulting in an equivalent two-dimensional measurement process of the form

$$\begin{bmatrix} dz_1(t) \\ dz_2(t) \end{bmatrix} = \begin{bmatrix} \cos x_1(t) \\ \sin x_1(t) \end{bmatrix} dt + \begin{bmatrix} dv_1(t) \\ dv_2(t) \end{bmatrix}, \quad (1)$$

where  $A$  has been taken as unity without loss of generality, and the noise has been replaced by a vector of mutually independent quantities.

The first-order phase process studied in [7] consisted of Brownian motion with increment of length  $h$  having variance  $qh$ . In this paper we describe a study of a second order phase process involving the integral of Brownian motion, expressed as

$$\begin{bmatrix} dx_1 \\ dx_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} dt + \begin{bmatrix} 0 \\ 1 \end{bmatrix} d\beta_t \quad (2)$$

We will retain the same measurement model (1) and let the noises  $v_1$  and  $v_2$  be independent Brownian motions with paths of length  $h$  having variance  $rh$ .

The familiar technique for tracking such phase processes involves the application of the phase-locked loop, studied thoroughly by Viterbi [8]. The phase-locked loop is a very ingenious nonlinear estimator, capable of near-optimal performance in good signal/noise environments. The steady-state behavior of the loop is identical to that of the so-called "linearized" or "extended" Kalman-Bucy filter for nonlinear systems, however, as was illustrated in [7]. Thus, the phase-locked loop is an excellent example of a very successful extended Kalman-Bucy filter. In the following section we discuss the stationary behavior of the linearized filter and the consequences of time discretization of the problem. Then in the subsequent section we will recast the discrete-time solution in terms of optimal nonlinear estimation, thereby setting the stage for a description of the numerical experiments.

#### B. The Linearized Kalman-Bucy Filter

Equations for the standard, continuous, linearized Kalman-Bucy filter are reproduced in Table 1 for the above phase-estimation problem. The measurement function is linearized about the current estimate  $\hat{x}_1(t)$  of the phase. The approximate filter attempts to track the mean of the conditional phase density, which, of course, is the minimum mean-squared error estimate. As a result, the loop estimate takes on all real values, whereas the original problem is only

Table 1. Summary of Continuous  
Linearized Kalman-Bucy Filter

185

Phase Process Model

$$\underline{dx} = \underline{F}\underline{x}dt + \underline{G}d\beta, \quad \underline{x}(0) \sim N[\underline{o}, \Sigma(0)]$$

$$\underline{F} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \underline{G} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad E(\beta_t - \beta_0)^2 = \int_0^t q \, ds.$$

Observation Model

$$\underline{dz} = \underline{h}(\underline{x})dt + \underline{dv}, \quad E(\underline{v}_t - \underline{v}_0)(\underline{v}_t - \underline{v}_0)' = \int_0^t \underline{R} \, ds,$$

$$\underline{R} = \begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix}, \quad \underline{h}(\underline{x}) = \begin{bmatrix} \cos x_1 \\ \sin x_1 \end{bmatrix}.$$

Filter Model

$$\underline{\hat{dx}} = (\underline{F} - \underline{K}\underline{H})\underline{\hat{x}}dt + \underline{K}(\underline{dz} - \underline{d\hat{z}} + \underline{H}\underline{\hat{x}}) \quad (4)$$

$$= \underline{F}\underline{\hat{x}}dt + \underline{K}\underline{dz},$$

$$\underline{\hat{dz}} = \underline{h}(\underline{\hat{x}}), \quad \underline{H} = \nabla_{\underline{x}} \underline{h}(\underline{\hat{x}}) = \begin{bmatrix} -\sin \hat{x}_1 & 0 \\ \cos \hat{x}_1 & 0 \end{bmatrix},$$

$$\underline{K} = \underline{P}\underline{H}'\underline{R}^{-1} = \begin{bmatrix} \frac{-P_{11}}{r} \sin \hat{x}_1 & \frac{P_{11}}{r} \cos \hat{x}_1 \\ \frac{-P_{12}}{r} \sin \hat{x}_1 & \frac{P_{12}}{r} \cos \hat{x}_1 \end{bmatrix},$$

$$\dot{\underline{P}} = \underline{F}\underline{P} + \underline{P}\underline{F}' - \underline{P}\underline{H}'\underline{R}^{-1}\underline{H}\underline{P} + \underline{Q} \quad (5)$$

$$= \begin{bmatrix} \dot{P}_{11} & \dot{P}_{12} \\ \dot{P}_{12} & \dot{P}_{22} \end{bmatrix} = \begin{bmatrix} 2P_{12} - \frac{P_{11}^2}{r} & P_{22} - \frac{P_{11}P_{22}}{r} \\ P_{22} - \frac{P_{11}P_{12}}{r} & q - \frac{P_{12}^2}{r} \end{bmatrix}$$

Table 1. Continued.

Equilibrium Solution

$$P = \begin{bmatrix} \sqrt{2} r^{3/4} q^{1/2} & r^{1/2} q^{1/2} \\ r^{1/2} q^{1/2} & \sqrt{2} r^{1/4} q^{3/4} \end{bmatrix} = r^{1/2} q^{1/2} \begin{bmatrix} \tau & 1 \\ 1 & \frac{2}{\tau} \end{bmatrix} \quad (6)$$

Filter Time Constant

$$\tau = \sqrt{2} \left( \frac{r}{q} \right)^{1/4}$$

observable modulo  $2\pi$ . Accordingly, it makes sense to consider the modulus of the error in the interval  $[-\pi, \pi)$ , or equivalently to take  $E[(e+\pi) \bmod 2\pi - \pi]^2$  as an error criterion. Naturally if the signal to noise ratio is high enough we would expect the minimum of the mean-modulo- $2\pi$ -squared error to be essentially equivalent to mean-squared error, but for higher noise situations, the modulation of the error would tend to bound the maximum mean-modulo- $2\pi$ -squared error (since the worst case will be a uniform error density on  $[-\pi, \pi)$ ). In Section C below we will discuss a nonlinear estimate designed to be defined only on  $[-\pi, \pi)$ .

In order to implement the nonlinear filter on a digital computer we will need to discretize time by some interval  $\Delta$ . The selection of  $\Delta$  will be made so as to assure essentially the same steady-state performance of the continuous and discrete phase-locked loops. Accordingly, we will now evaluate the steady-state solution of the matrix Riccati-equation for the linearized filter, which we know is approximately equal to the error variance for low-noise applications. If we make the definitions

$$\begin{aligned} F &\triangleq \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, & Q &\triangleq \begin{bmatrix} 0 & 0 \\ 0 & q \end{bmatrix}, \\ H &\triangleq \begin{bmatrix} -\sin \hat{x}_1 & 0 \\ \cos \hat{x}_1 & 0 \end{bmatrix}, & R &\triangleq \begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix}, \end{aligned} \quad (3)$$

then the linearized filter satisfies the equations

$$d\hat{\underline{x}} = (F - KH)\hat{\underline{x}}dt + K(d\underline{z} - \hat{\underline{z}} + H\hat{\underline{x}}dt) \quad (4)$$

where

$$d\hat{\underline{z}} = \begin{bmatrix} \cos \hat{x}_1 \\ \sin \hat{x}_1 \end{bmatrix} dt, \quad \hat{\underline{x}}(0) = 0$$

and

$$K = PH'R^{-1},$$

where  $P$  satisfies

$$\frac{dP}{dt} = FP + PF' - PH'R^{-1}HP + Q, \quad (5)$$

with  $P(0) = \Sigma$ . This system is diagrammed in Fig. 1. We may set the derivative in (5) to zero as a necessary condition for steady-state and algebraically determine the possible steady-state solutions.

Accordingly, we obtain

$$\begin{aligned} P &= \begin{bmatrix} \sqrt{2} r^{3/4} q^{1/4} & r^{1/2} q^{1/2} \\ r^{1/2} q^{1/2} & \sqrt{2} r^{1/4} q^{3/4} \end{bmatrix} \\ &= \sqrt{rq} \begin{bmatrix} \sqrt{2} \left(\frac{r}{q}\right)^{1/4} & 1 \\ 1 & \sqrt{2} \left(\frac{q}{r}\right)^{1/4} \end{bmatrix} \end{aligned} \quad (6)$$

$$= \sqrt{rq} \begin{bmatrix} \tau & 1 \\ 1 & \frac{2}{\tau} \end{bmatrix}, \quad (7)$$

where  $\tau = \sqrt{2} \left(\frac{r}{q}\right)^{1/4}$ , the filter time constant, is obtained by studying the  $\hat{\underline{x}}$  equation (4) with (6) substituted for  $P$  as follows: The homogeneous part of (4) is rewritten as

$$\begin{aligned} d\hat{\underline{x}} &= (F - KH) \hat{\underline{x}} dt \\ &= (F - PH'R^{-1}H) \hat{\underline{x}} \\ &= \begin{bmatrix} -\frac{1}{r} \sqrt{2} r^{3/2} q^{1/2} & 1 \\ -\frac{1}{r} \sqrt{rq} & 0 \end{bmatrix} \hat{\underline{x}} dt. \end{aligned} \quad (8)$$

The eigenvalues of the matrix in (8) are the solutions  $\lambda$  to

$$\lambda^2 + \frac{1}{r} \sqrt{2} r^{3/2} q^{1/2} \lambda + \frac{1}{r} \sqrt{rq} = 0, \quad (9)$$

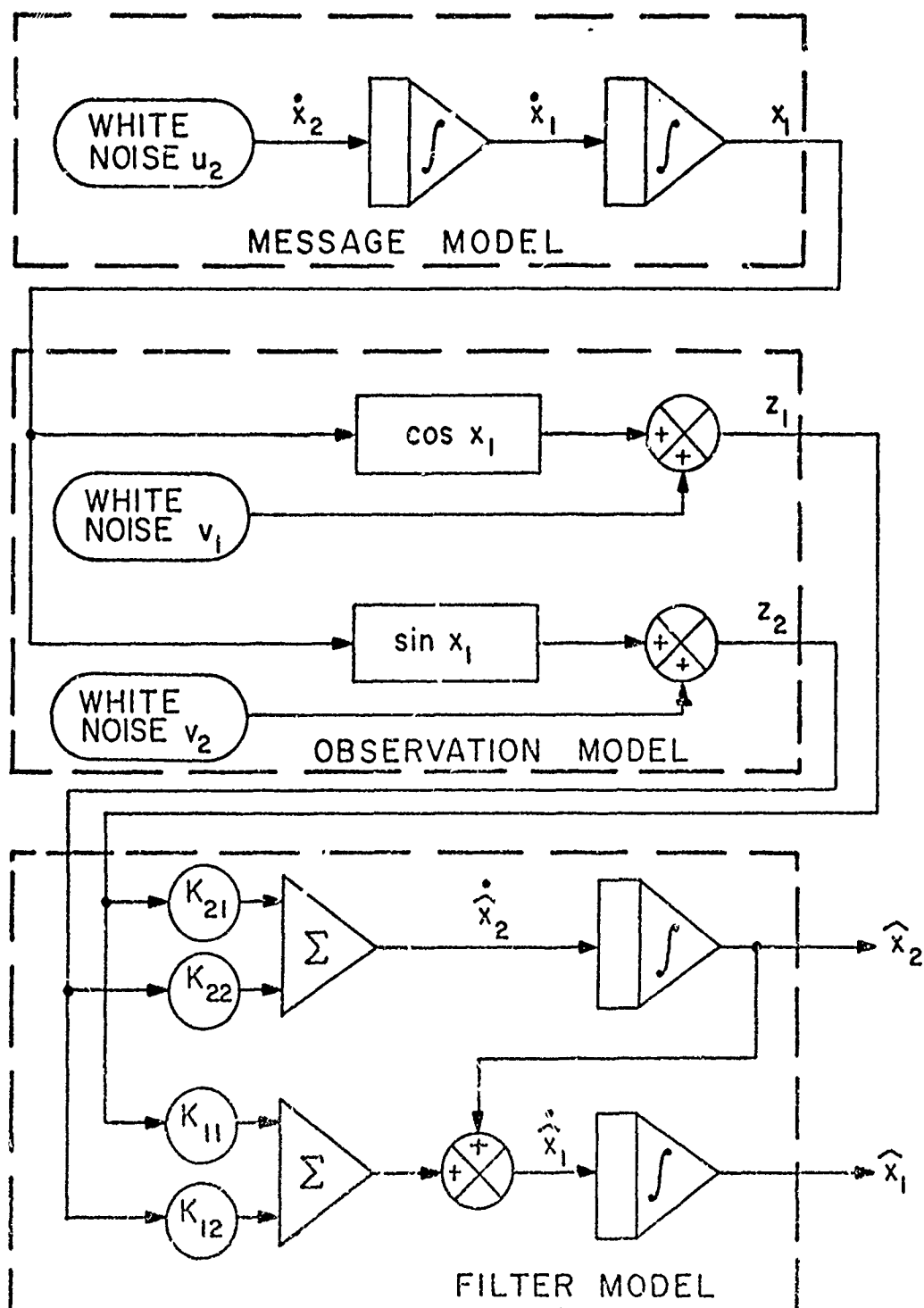


Fig. 1. Block Diagram of Linearized Phase Estimation

resulting in  $\lambda = \frac{1}{\sqrt{2}} \left(\frac{q}{r}\right)^{1/4} (-1 \pm i)$ , or, using the definition for  $\tau$  in (7),

$$\lambda = \frac{1}{\tau} (-1 \pm i). \quad (10)$$

Thus the solutions to (8) are of the form

$$\hat{x}_1 = C_0 e^{-\frac{t}{\tau}} \cos\left(\frac{t}{\tau} + C_1\right),$$

so that  $\tau$  is indeed a time constant. From (4) and (8) we may write the steady-state differential equations for the extended Kalman-Bucy filter (i.e., the phase-locked loop) as

$$d\hat{x} = F \hat{x} dt + K (dz - d\hat{z}), \quad (11)$$

or, equivalently,

$$d\hat{x}_1 = \hat{x}_2 dt + \frac{2}{\tau} (-\sin \hat{x}_1 dz_1 + \cos \hat{x}_1 dz_2),$$

$$\text{and} \quad d\hat{x}_2 = \frac{2}{\tau} (\sin \hat{x}_1 dz_1 + \cos \hat{x}_1 dz_2). \quad (12)$$

Refer to Table 1 for a summary of above results. The corresponding results for the discrete-time filter, obtained by Hecht [6], are summarized in Table 2. If we are interested in simulating a filter which behaves substantially the same as the continuous filter then we must choose the sampling interval carefully so that it is as large as possible subject to the constraint that the discrete covariance in steady-state is within an acceptable tolerance of the continuous covariance.

The steady-state discrete prediction covariance  $S$  and the filtering covariance  $P_d$  must converge to the continuous  $P$  of (6) in the limit as  $\Delta \rightarrow 0$ . Accordingly, we write  $S$  as a function of  $P$  and  $\Delta/\tau$  as

$$S_{11}\left(\frac{\Delta}{\tau}\right) = P_{11} \frac{\tau}{2\Delta} \left(\frac{1}{\alpha_0} - 1\right), \quad (13)$$

Table 2. Summary of Discrete  
Linearized Kalman-Bucy Filter

Phase Process Model

$$\underline{x}(n\Delta) = \Phi[n\Delta, (n-1)\Delta] \underline{x}[(n-1)\Delta] + \Gamma \underline{u}_d(n\Delta) ,$$

$$\Phi[n\Delta, (n-1)\Delta] = e^{F\Delta} \approx \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix} \triangleq \Phi$$

$$E[u_d^2(n\Delta)] = q_d = \frac{q}{\Delta} , \quad n\Delta = t$$

$$\Gamma = \int_{(n-1)\Delta}^{n\Delta} \Phi[n\Delta, \mu] G d\mu \approx G\Delta = \begin{bmatrix} 0 \\ \Delta \end{bmatrix}$$

Observation Model

$$\underline{z}(n\Delta) = \underline{h}[\underline{x}(n\Delta)] + \underline{v}_d(n\Delta)$$

$$\underline{h}[\underline{x}(n\Delta)] = \begin{bmatrix} \cos x_1(n\Delta) \\ \sin x_1(n\Delta) \end{bmatrix}$$

$$E[\underline{v}_d(n\Delta) \underline{v}_d'(n\Delta)] = R_d = \begin{bmatrix} \frac{r}{\Delta} & 0 \\ 0 & \frac{r}{\Delta} \end{bmatrix}$$

Table 2. Continued.

Predictor Model

$$\hat{\underline{x}}(n|n-1) = \Phi \hat{\underline{x}}(n|n) \quad \hat{\underline{x}}(0|0) = \underline{0}$$

$$S(n) = \Phi P_d(n-1) \Phi' + \Gamma Q_d \Gamma'$$

$$= \begin{bmatrix} S_{11}(n) & S_{12}(n) \\ S_{12}(n) & S_{22}(n) \end{bmatrix}$$

Filter Model

$$\hat{\underline{x}}(n|n) = \hat{\underline{x}}(n|n-1) + A(n) [\underline{z}(n) - \hat{\underline{z}}(n|n-1)]$$

$$= \hat{\underline{x}}(n|n-1) + A(n) \underline{z}(n) ,$$

$$\hat{\underline{z}}(n|n-1) = h[\hat{\underline{x}}(n|n-1)] ,$$

$$A(n) = S(n)H' [H S(n)H' + R_d]^{-1}$$

$$= \frac{1}{S_{11} + r_d} \begin{bmatrix} -S_{11}(n) \sin \hat{x}_1(n|n-1) & S_{11}(n) \cos \hat{x}_1(n|n-1) \\ -S_{12}(n) \sin \hat{x}_1(n|n-1) & S_{12}(n) \cos \hat{x}_1(n|n-1) \end{bmatrix}$$

$$P_d(n) = [I - A(n)H] S(n) [I - A(n)H]' + A(n) R_d A(n)$$

$$= \frac{1}{S_{11} + r_d} \begin{bmatrix} S_{11}(n) & S_{12}(n) \\ S_{12}(n) & S_{22}(n) \left( \frac{S_{11}(n) + r_d}{r_d} \right) - \frac{S_{12}^2(n)}{r_d} \end{bmatrix}$$

Table 2. Continued.

Equilibrium Solution

$$S_{11}(n+1) = \frac{r_d [S_{11}(n) + 2 S_{12}(n)\Delta] - S_{12}^2(n)\Delta^2}{S_{11}(n) + r_d} + S_{22}(n)^2$$

$$S_{12}(n+1) = \frac{S_{12}(n)[r_d - S_{12}(n)\Delta]}{S_{11}(n) + r_d} + S_{22}(n)\Delta$$

$$S_{22}(n+1) = \frac{-S_{12}^2(n)}{S_{11}(n) + r_d} + S_{22}(n) + \Delta^2 q_d$$

Setting  $S(n+1) = S(n) = S$ , using a discrete form of the Bass-Roth Theorem [1],

$$S = \begin{bmatrix} r_d \left( \frac{1}{\alpha_0} - 1 \right) & \frac{\Delta \sqrt{q_d r_d}}{\sqrt{\alpha_0}} \\ \frac{\Delta \sqrt{q_d r_d}}{\sqrt{\alpha_0}} & \sqrt{q_d r_d} \left( \frac{1}{\sqrt{\alpha_0}} - \sqrt{\alpha_0} \right) - \Delta^2 q_d \end{bmatrix}$$

$$= \begin{bmatrix} \frac{r}{\Delta} \left( \frac{1}{\alpha_0} - 1 \right) & \frac{\sqrt{q_d r}}{\sqrt{\alpha_0}} \\ \frac{\sqrt{q_d r}}{\sqrt{\alpha_0}} & \frac{\sqrt{q_d r}}{\Delta} \left( \frac{1}{\sqrt{\alpha_0}} - \sqrt{\alpha_0} \right) + \Delta q \end{bmatrix},$$

with

$$\alpha_0 = 1 + \frac{p}{4} + \frac{m^{1/2}}{2} - \frac{1}{2} \left[ \frac{p^2}{2} + 6p + (4+p)m^{1/2} \right]^{1/2},$$

$$m = \frac{\rho}{2}^2 + 4\rho$$

$$\rho = \frac{q_d}{r_d} \Delta^4$$

$$S_{12}\left(\frac{\Delta}{\tau}\right) = P_{12} \frac{1}{\sqrt{\alpha_o}}, \quad (14)$$

$$S_{22}\left(\frac{\Delta}{\tau}\right) = P_{22} \left[ \frac{\tau}{2\Delta} \left( \frac{1}{\sqrt{\alpha_o}} - \sqrt{\alpha_o} \right) + \frac{\Delta}{\tau} \right], \quad (15)$$

where  $\alpha_o$  is a function of  $\frac{\Delta}{\tau}$ , since the  $\rho$  of Table 2 may be written

$$\rho\left(\frac{\Delta}{\tau}\right) = 4\left(\frac{\Delta}{\tau}\right)^4. \quad (16)$$

The filtering steady state  $P_d$  may be similarly expressed as

$$P_{d11}\left(\frac{\Delta}{\tau}\right) = \alpha_o S_{11}\left(\frac{\Delta}{\tau}\right) = P_{11} \frac{\tau}{2\Delta} (1 - \alpha_o), \quad (17)$$

$$P_{d12}\left(\frac{\Delta}{\tau}\right) = \alpha_o S_{12}\left(\frac{\Delta}{\tau}\right) = P_{12} \sqrt{\alpha_o}, \quad (18)$$

$$\begin{aligned} P_{d22}\left(\frac{\Delta}{\tau}\right) &= S_{22}\left(\frac{\Delta}{\tau}\right) - \frac{S_{12}^2\left(\frac{\Delta}{\tau}\right)}{S_{11}\left(\frac{\Delta}{\tau}\right) + r_d} \\ &= P_{22} \left[ \frac{\tau}{2\Delta} \left( \frac{1}{\sqrt{\alpha_o}} - \sqrt{\alpha_o} \right) + \frac{\Delta}{\tau} \right] - \frac{P_{12}^2}{P_{11}} (1 - \alpha_o) \end{aligned} \quad (19)$$

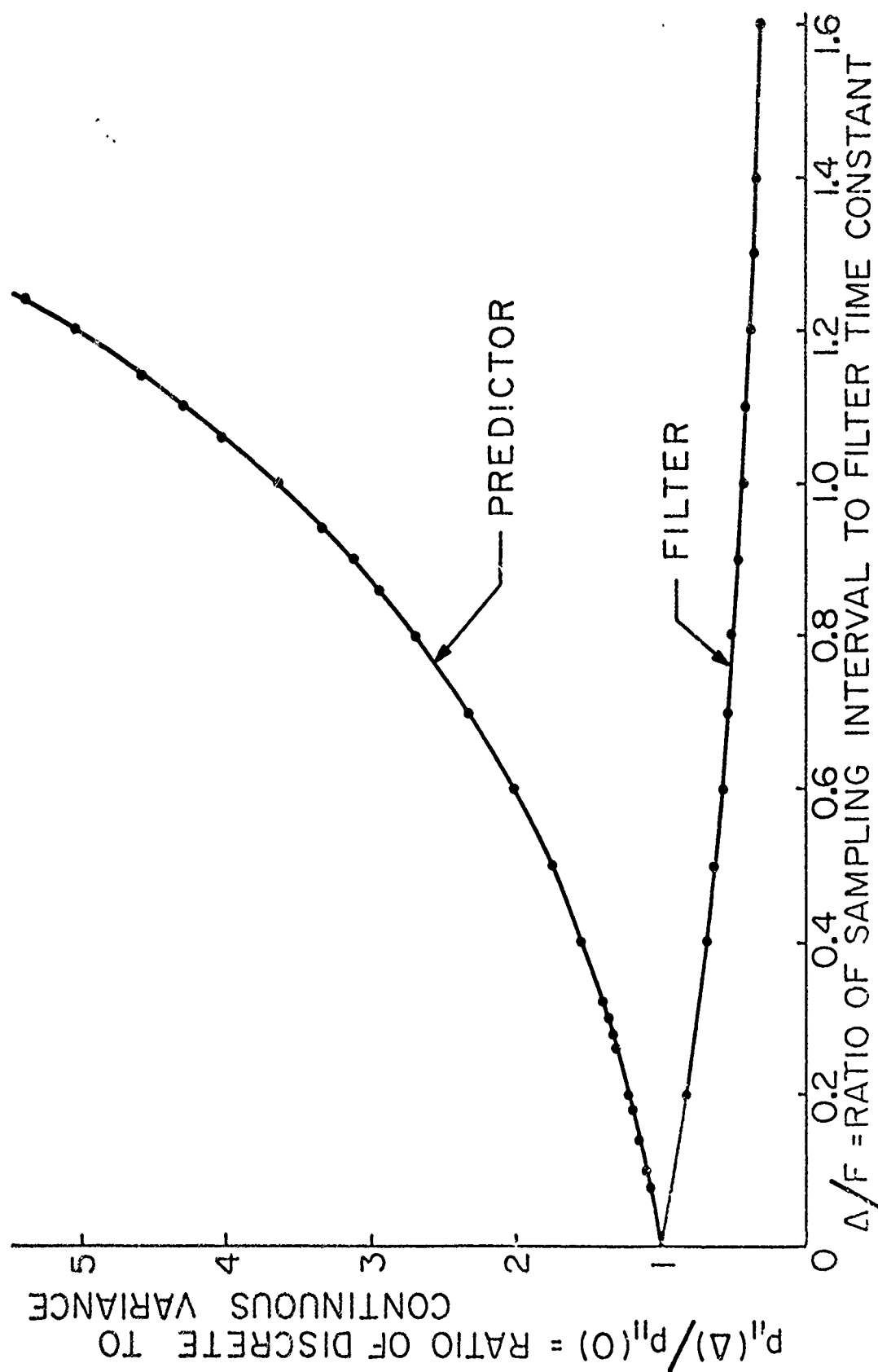
The relationship between  $\Delta/\tau$  and the discrete variances is illustrated in Figs. 2-4. If we choose  $\Delta/\tau = 0.1$  and evaluate (13) and (17), we find

$$\frac{S_{11}(0.1)}{P_{11}} = 1.108028,$$

while

$$\frac{P_{d11}(0.1)}{P_{11}} = 0.9070263,$$

or  $S_{11}(0.1) - P_{d11}(0.1) \approx 0.2 P_{11}$ . Thus we suffer a 10% change in the steady-state covariance by taking 10 samples per time constant. In any case we will compare all discrete filters to each other, and we can assume that all results lie within 10% of their continuous limits.

Fig. 2. Discrete  $p_{11}$  Error Variance

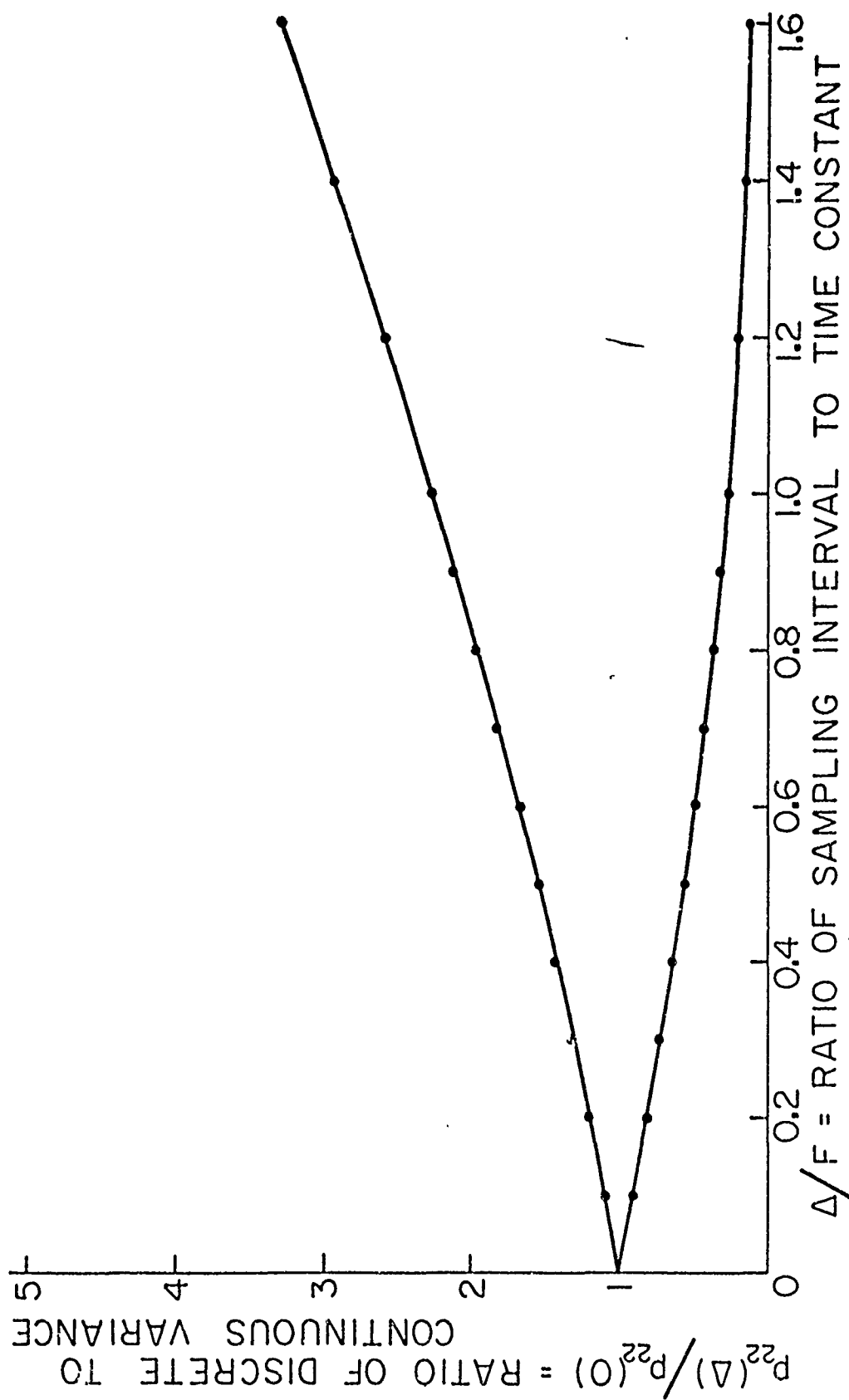


Fig. 3. Discrete  $P_{22}$  Error Variance

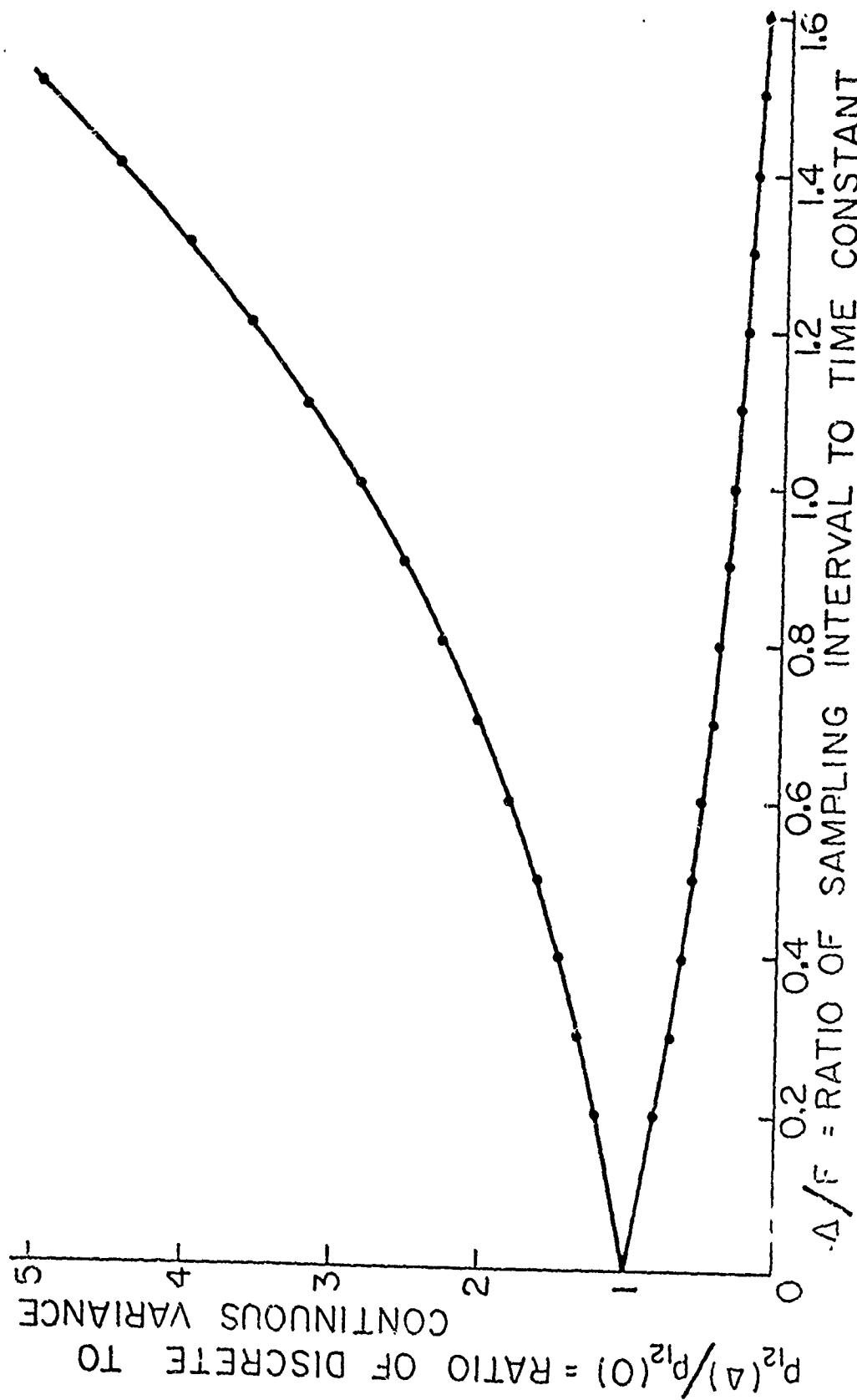


Fig. 4. Discrete  $P_{12}$  Error Variance

### C. Application of Nonlinear Filtering

The application of Bayes-Law filtering, developed by Bucy and Senne [5], requires special considerations for each special case. In the problem of this paper, for example, the dimension of the driving noise is one less than that of the state vector, a situation for which Bucy and Senne indicate there results in a computational simplification for Bayes Law implementation. We will explore that claim here for the second order phase process.

At the outset consider a similar but unrealistic phase process with two driving noises  $u_1$  and  $u_2$  so that (in discrete time)

$$\underline{x}(n+1) = \phi \underline{x}(n) + \Gamma \underline{u}(n), \quad (20)$$

where

$$\Gamma = \begin{bmatrix} \epsilon \Delta & 0 \\ 0 & \Delta \end{bmatrix}, \quad \underline{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix},$$

and the remaining quantities, as well as the sensor model remain the same as described in Table 2. Now if we let the variance of  $u_2$  and  $u_1$  be  $q/\Delta$  and take the limit as  $\epsilon \rightarrow 0$  then the process (20) will be identical to the one described in Table 2. For any finite  $\epsilon$ , however, the probability density function of the driving terms  $\Gamma \underline{u}$  will take the form

$$P_{\Gamma \underline{u}}(\underline{\xi}) = \frac{1}{2\pi \det[A]} \exp \left\{ -\frac{1}{2} \|\underline{\xi}\|_{A^{-1}}^2 \right\}, \quad (21)$$

where

$$A \triangleq \Gamma Q \Gamma^T = \begin{bmatrix} \epsilon^2 \Delta q & 0 \\ 0 & \Delta q \end{bmatrix}.$$

Similarly, the density of the observation noise will be denoted

$P_v(\underline{\xi})$ , which will be gaussian with covariance  $R$ .

If we review the Bayes representation theorem solution to the discrete filtering problem [4], we can determine that

$$J_{n+1|n}(\underline{y}) = \frac{1}{K_n} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P_{\Gamma u}(\underline{y} - \phi \underline{x}) J_{n|n}(\underline{x}) dx_1 dx_2, \quad (22)$$

and

$$J_{n|n}(\underline{y}) = \frac{1}{K_n} P_v \left[ z_n - h(\underline{y}) \right] J_{n|n-1}(\underline{y}) \quad (23)$$

Now if we take the limit as  $\epsilon \rightarrow 0$  in (22) we obtain

$$\begin{aligned} J_{n+1|n}(\underline{y}) &= \frac{C}{K_n} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta[y_1 - x_2 \Delta - x_1] \exp \left\{ -\frac{1}{2q\Delta} (y_2 - x_2)^2 \right\} J_{n|n}(\underline{x}) dx_1 dx_2 \\ &= \frac{C}{K_n} \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2q\Delta} (y_2 - x_2)^2 \right\} J_{n|n} \left( \begin{matrix} y_1 - x_2 \Delta \\ x_2 \end{matrix} \right) dx_2, \end{aligned} \quad (24)$$

where  $C = (2\pi)^{-1/2} (q\Delta)^{-1/2}$ . In other words the Bayes integral reduces to an integral over a subspace of the state space with dimension equal to that of the driving noise vector (equals one in this case). The complete Bayes recursion may now be written

$$J_{n+1|n}(\underline{y}) = C_1 \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2q\Delta} (y_2 - x_2)^2 \right\} J_{n|n} \left( \begin{matrix} y_1 - x_2 \Delta \\ x_2 \end{matrix} \right) dx_2, \quad (25)$$

$$J_{n|n}(\underline{y}) = C_2 \exp \left\{ -\frac{\Delta}{2r} \left[ (z_1 - \cos y_1)^2 + (z_2 - \sin y_1)^2 \right] \right\} J_{n|n-1}(\underline{y}), \quad (26)$$

where both  $C_1$  and  $C_2$  are normalizing constants.

Although there can be no doubt that a significant simplification has resulted from reducing the number of integrations required for the Bayes integral computation, it must be pointed out that the arguments of  $J_{n|n}$  in (25) must be determined to lie in a particular subspace of  $R^2$ . If the densities have been stored only at a finite number of points

then some form of interpolation of the densities will be necessary, resulting in some overhead, so that the computation is not exactly equivalent to a scalar problem.

Another problem of concern to this particular application is the domain of the conditional probability densities. If we are only interested in modulo- $2\pi$  errors for the reasons stated earlier then there is no loss of generality to map all modulo- $2\pi$  intervals of the conditional densities back into the  $[-\pi, \pi)$  interval in the  $x_1$  coordinate and sum up the individual contributions. Similarly, because we are computing the phase in discrete-time with sample time  $\Delta$ , we observe that phase-rate contributions outside the interval  $[-\frac{\pi}{\Delta}, \frac{\pi}{\Delta})$  will have the same effect on the next phase as their modulo- $\frac{2\pi}{\Delta}$  component. Accordingly, we may map the phase-rate  $x_2$  components of the conditional densities back into  $[-\frac{\pi}{\Delta}, \frac{\pi}{\Delta})$  for the same reason as above. The combination of the two mappings results in an equivalence between the "cyclic" state space and a torus (see Fig. 5).

Next we consider what simplifications arise for the Bayes integral update (25)-(26) as a consequence of the cyclic mapping of the state space. We begin by combining (25) and (26) into a single equation representing the filter update, and absorb all non state-dependent terms of the quadratic exponent expansions into a single normalizing constant  $C_0$ , giving

$$J_{n+1|n+1} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = S_{n+1}(y_1) \int_{-\infty}^{\infty} \exp \left\{ -\frac{(y_2 - \mu)^2}{2q\Delta} \right\} J_{n|n} \begin{pmatrix} y_1 - \mu\Delta \\ \mu \end{pmatrix} d\mu, \quad (27)$$

where

$$S_{n+1}(y_1) \triangleq C_0 \exp \left\{ \frac{Z_1(n+1) \cos y_1 + Z_2(n+1) \sin y_1}{r/\Delta} \right\}. \quad (28)$$

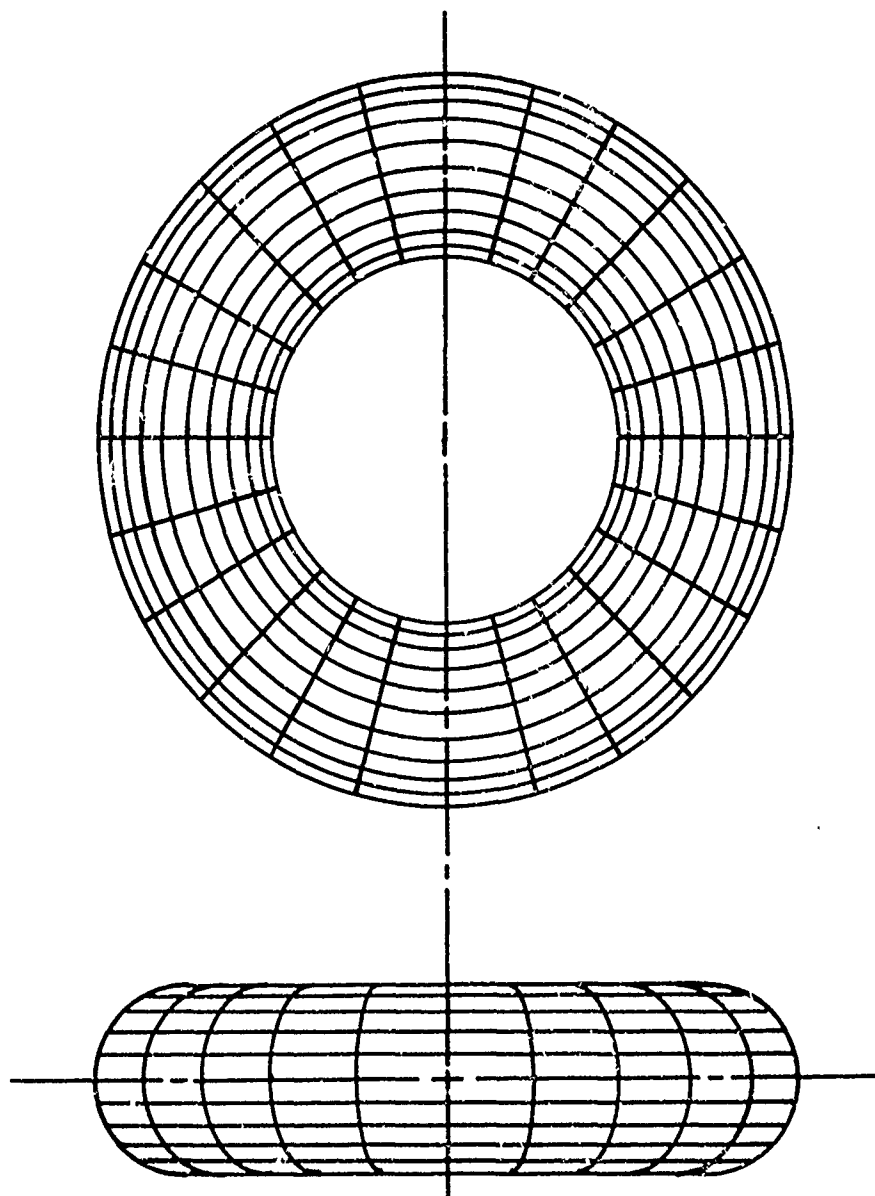


Fig. 5. Torus Interpretation of Doubly Cyclic  
State Space

If we now modulate the density  $J_{n|n}$  as described above, we obtain

$$\tilde{J}_{n|n} \left( \begin{matrix} \sigma \\ \tau \end{matrix} \right) \triangleq \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} J_{n|n} \left( \begin{matrix} \sigma + 2\pi k \\ \tau + \frac{2\pi l}{\Delta} \end{matrix} \right), \quad (29)$$

with  $-\pi \leq \sigma < \pi$ , and  $\frac{-\pi}{\Delta} \leq \tau < \frac{\pi}{\Delta}$ . Finally we use the definition (29) on  $J_{n+1|n+1}$ , and substitute (27), resulting in the following manipulations [2]:

$$\tilde{J}_{n+1|n+1} \left( \begin{matrix} \sigma \\ \tau \end{matrix} \right) = \sum_k \sum_l S_{n+1}(\sigma + 2\pi k) \int_{-\infty}^{\infty} \exp \left\{ -\frac{(\tau + 2\pi l/\Delta - \mu)^2}{2q\Delta} \right\} J_{n|n} \left( \begin{matrix} \sigma - \mu\Delta + 2\pi k \\ \mu \end{matrix} \right) d\mu \quad (30)$$

$$= S_{n+1}(\sigma) \sum_k \sum_l \sum_m \int_{\frac{-\pi + 2\pi m}{\Delta}}^{\frac{\pi + 2\pi m}{\Delta}} \exp \left\{ -\frac{(\tau + 2\pi l/\Delta - \mu)^2}{2q\Delta} \right\} J_{n|n} \left( \begin{matrix} \sigma - \mu\Delta + 2\pi k \\ \mu \end{matrix} \right) d\mu$$

$$(\text{let } \xi = \mu - \frac{2\pi m}{\Delta})$$

$$= S_{n+1}(\sigma) \sum_k \sum_l \sum_m \int_{-\frac{\pi}{\Delta}}^{\frac{\pi}{\Delta}} \exp \left\{ -\frac{[\tau - \xi + 2\pi(1-m)/\Delta]^2}{2q\Delta} \right\} J_{n|n} \left( \begin{matrix} \sigma - \xi\Delta + 2\pi(k-m) \\ \xi + \frac{2\pi m}{\Delta} \end{matrix} \right) d\xi$$

$$(\text{let } i = 1-m, j = 1-i)$$

$$= S_{n+1}(\sigma) \sum_i \sum_j \sum_k \int_{-\frac{\pi}{\Delta}}^{\frac{\pi}{\Delta}} \exp \left\{ -\frac{(\tau - \xi + 2\pi i/\Delta)^2}{2q\Delta} \right\} J_{n|n} \left( \begin{matrix} \sigma - \xi\Delta + 2\pi(k-j) \\ \xi + \frac{2\pi j}{\Delta} \end{matrix} \right) d\xi$$

(Fubini's theorem to interchange  $\sum$  and  $\int$ )

$$= S_{n+1}(\sigma) \int_{-\frac{\pi}{\Delta}}^{\frac{\pi}{\Delta}} a(\tau - \xi) \tilde{J}_{n|n} \left( \begin{matrix} \sigma - \xi\Delta \\ \xi \end{matrix} \right) d\xi, \quad (31)$$

where

$$a(\tau-\xi) \triangleq \sum_{i=-\infty}^{\infty} \exp \left\{ - \frac{(\tau-\xi+2\pi i/\Delta)^2}{2q\Delta} \right\} \quad (32)$$

The result (31) represents the exact recurrence relation required for the cyclic conditional density. The individual terms of  $a(\cdot)$  will drop rapidly on either side of  $\tau-\xi$ , depending only on the variance  $q\Delta$ .

Having constructed a density function updating formula for the phase estimation problem, the question remains - what form shall the phase estimate itself take? It is clear that the conditional mean (the goal of the phase-locked loop) is an admissible candidate. The cost criterion that is minimized by the conditional mean, however, is the mean squared error. It is not obvious that mean squared error is the best criterion for choosing estimates modulo  $2\pi$ . In fact, a periodic cost function of the form

$$L(e) = 2(1 - \cos e) \quad (33)$$

might be more appropriate than  $e^2$  if only modulo- $2\pi$ -errors are important. The cyclic loss (33) looks like  $e^2$  for small  $e$  and  $(e-2k\pi)^2$  for  $e$  close to  $2k\pi$  for all  $k$ . Moreover, we may easily show that the estimate  $x_n^*$  which minimizes

$$E[L(x_n^* - x_n^1) | Z_n] = \int_{-\pi}^{\pi} \int_{-\frac{\pi}{\Delta}}^{\frac{\pi}{\Delta}} L(\tau - x_n^*) \tilde{J}_n(\tau, \sigma) d\sigma d\tau \quad (34)$$

is given by

$$x_n^* = \tan^{-1} \left\{ E[\sin(x_n) | Z_n] / E[\cos(x_n) | Z_n] \right\} \quad (35)$$

For a proof of (35) see [7].

Of course the cyclic loss may not be the only desirable loss function for the phase demodulator. But many other proposed criteria would be

minimized by appropriate functions of the conditional phase densities, resulting in extremely great flexibility for the experiment designer. Moreover, it is seen that the conditional expected loss (34) as well as the densities themselves are computable in addition to the estimates so that a considerable amount of quantitative information is available to provide a realistic assessment of the quality of the estimates, regardless of the cost criterion employed. No such information is provided by the phase-locked loop - especially after steady-state is attained.

## References

- [1] M. Abramowitz and I.A. Stegun, Handbook of Mathematical Functions, Dover, New York, 1965.
- [2] R.S. Bucy, "Building and Evaluating Non-Linear Filters," To appear Proc. Symp. on Appl. Math.; Stochastic Diff. Eqns., Am. Math. Soc., April 1972.
- [3] R.S. Bucy, C. Hecht, and K.D. Senne, "Optimal Phase Demodulation via Discrete Nonlinear Filtering," Air Force Weapons Laboratory Computer Films No. 72-0401-01, April 1972.
- [4] R.S. Bucy and P.D. Joseph, Filtering for Stochastic Processes with Applications to Guidance, Wiley Interscience, New York, 1968.
- [5] R.S. Bucy and K.D. Senne, "Digital Synthesis of Nonlinear Filters," Automatica, 7 (1971), 278-298.
- [6] C. Hecht, "Synthesis and Realization of Nonlinear Filters," Ph.D. Dissertation, University of Southern California, January 1972.
- [7] A.J. Mallinckrodt, R.S. Bucy, and S.L. Cheng, "Final Report for a Design Study for an Optimal Nonlinear Receiver-Demodulator," NASA Contract NAS5-10789, Goddard Space Flight Center, Maryland, 1970.
- [8] A.J. Viterbi, Principles of Coherent Communication, McGraw Hill, New York, 1966.

### Appendix A. Numerical Experiments With The Phase-Locked Loop

In order to provide an accurate check on the value of the nonlinear filters, it was necessary to perform extensive Monte Carlo tests on the phase-locked loop (i.e. the steady-state linearized filter). Since the discrete phase-locked loop operates very fast (over 1000 estimates per second on the CDC 6600), it was possible to average estimates over 5000 sample paths of length 130 in 10 minutes. If three time constants are discarded (30 samples) in each sample path the resulting 100 estimates represent steady-state. If all of the steady-state errors were averaged this would lead to 500000 monte carlos of the steady-state error. On the other hand, since adjacent errors are correlated, the effective Monte Carlo length would better be set between  $N=50000$  (one estimate per time constant) and  $N=500000$  (every estimate included). Thus, we may determine the three-standard deviation confidence bands based on both values of  $N$ .

The independent parameter for the different phase-locked loop cases considered was  $p_{11}$ , the steady-state continuous Ricatti equation solution for the phase error variance.  $p_{11} = \sqrt{2r^{3/4}q^{1/4}}$  is shown by Viterbi [8] to be the inverse of the effective signal to noise ratio, or  $N/S$ . The initial condition for the matrix Ricatti equation can be taken to be its steady-state value\*, and the mean-squared error will

---

\*An alternative initial condition of  $p_{11}(0)=4p_{11}$  was used in Appendix B, where it was discovered that there is a significance dependence between the initial condition and the effective time constant of the loop. Thus the time required to reach steady-state from the large initial condition is frequently too long for an effective Monte Carlo analysis.

**Preceding page blank**

then only be a function of  $N/S$  and not of  $q$ . A value of 0.01 was chosen for  $q$ , and thus the value of  $r$  was taken to be  $(p_{11})^{4/3}/(2^{2/3}q^{1/3})$ .

Now if the phase-locked loop were linear with  $H=[-1 \ 0]$ , then  $p_{11}$  would be the steady-steady mean-squared error. An equivalent interpretation would be to consider  $p_{11}$  as the state-state mean-squared error of a filter operating on the linear measurements  $\underline{z} = \underline{H}x + \underline{v}$ . The latter interpretation verifies that in all cases the mean-squared error of the actual loop can be no lower than  $p_{11}$ . If we are interested in modulo- $2\pi$  errors we may convert the lower bound into mean-modulo- $2\pi$ -squared error by the equation

$$\begin{aligned}\sigma_m^2 &= \int_{-\infty}^{\infty} \left[ (e+\pi) \bmod 2\pi - \pi \right]^2 \exp\left[-\frac{e^2}{2p_{11}}\right] de \\ &= \int_{-\pi}^{\pi} e^2 \sum_{k=-\infty}^{\infty} \exp\left[-\frac{(e+2k\pi)^2}{2p_{11}}\right] de, \quad (A-1)\end{aligned}$$

where  $\sigma_m^2 \triangleq$  mean-modulo- $2\pi$ -squared error. The result  $\sigma_m^2$  of (A-1) is plotted in Fig. A-1 as well as the equivalent discrete filter result  $\sigma_{dm}^2$  (based on  $p_{d11}$ ).  $\sigma_m^2$  was determined by the Newton integration based on 10000 points in the interval  $[-\pi, \pi)$  and 15 standard deviations taken from the infinite sum.

The deviation the Monte-Carlo performance of the phase-locked loop from the ideal (linear) analysis can also be seen in Fig. A-1 to be insignificant below  $N/S = -10\text{db}$  or above  $+8\text{db}$ . The significant departure in the middle region is a result of the "cycle-slip" phenomena, whereby the unmodulated phase-error density begins to contain substantial probability in the secondary modes [8]. For

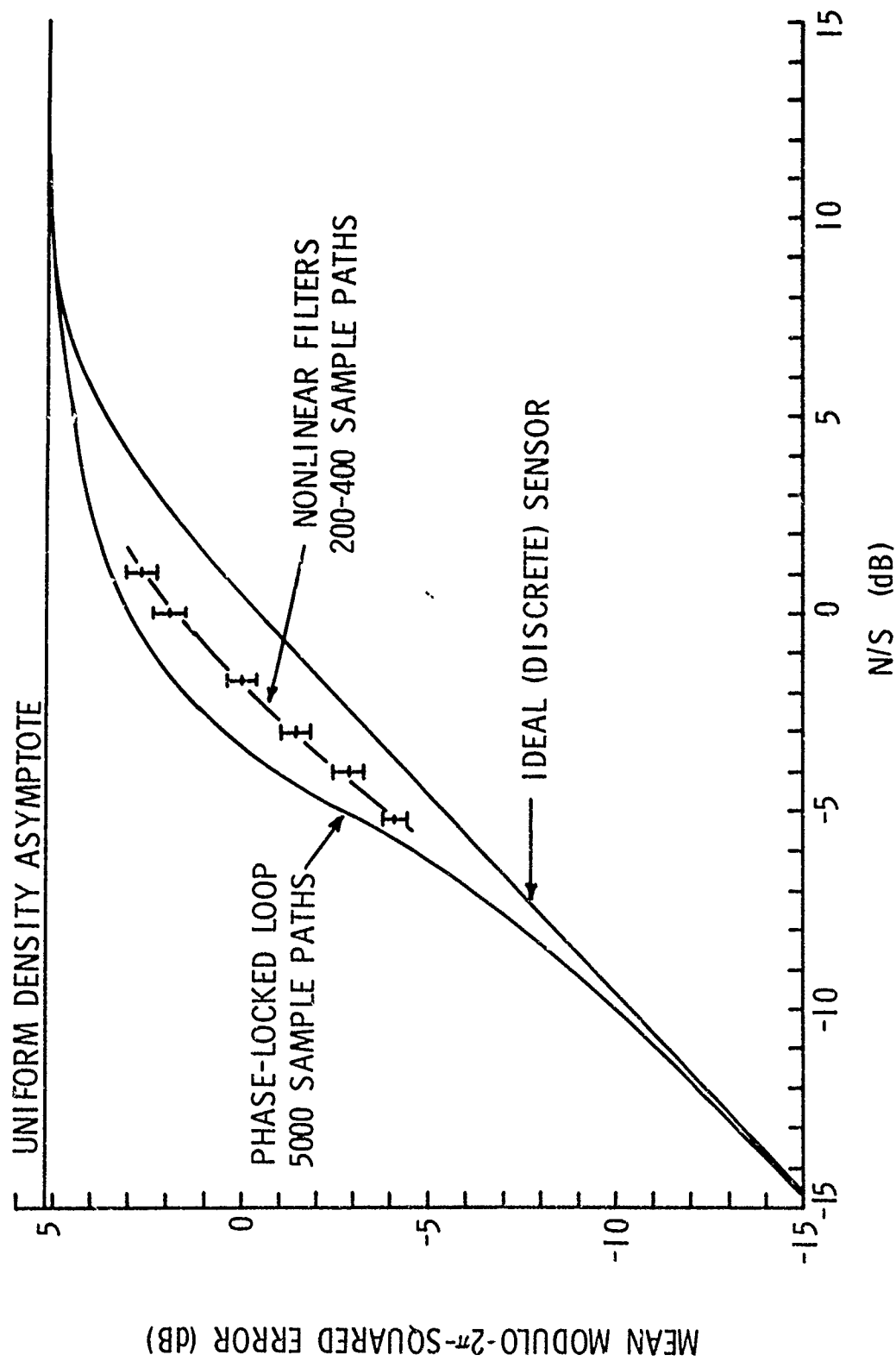


Fig. A-1. MSE Performance Summary

extremely high noise situations, however, modulation of the error density causes the asymptotic error density to become uniform on  $[-\pi, \pi)$ , resulting in an asymptotic variance of  $\pi^2/3 = 3.29$  (or 5.19 dB) for  $N/S$  large.

The point at which the -10dB departure between the two curves occurs is often referred to as "threshold", where unlock of the loop begins to cause problems. Almost all engineering modifications to the basic loop are designed for the purpose of extending the threshold. It is clear that all nonlinear filters must have modulo- $2\pi$  error variance in the region between the two curves (discrete phase-locked loop and discrete-ideal). Thus the problem of threshold extension is equivalent to reducing steady-state error variance, which attains the maximum potential improvement of about 4dB at -1.8dB N/S.

The confidence in the Monte Carlos of the phase-locked loop can be computed using the results from Chapter IV. Accordingly, we observed that a value for  $\rho_2 = \mu_4/\mu_2^2$  must be determined, since the three-standard deviation confidence bounds on variance depend on  $\rho_2$ . Thus our Monte Carlo simulations involved the calculation of an estimate  $\hat{\mu}_4$  of  $\mu_4$  and the estimate  $\hat{\mu}_2$  of  $\mu_2$ . Then we plotted the ratio  $\hat{\mu}_4/\hat{\mu}_2^2 = \hat{\rho}_2$  in Fig. A-2. From the figure we determine that a good upper bound on  $\rho_2$  is given by 5.4, which is achieved near -5.0 dB N/S. A summary of the estimate  $3\sigma$  confidence bands is given in Table A-1, where we observe that the maximum confidence band width is from  $\pm 0.038$  dB to  $\pm 0.120$  dB, depending on whether  $N$  was taken as 500000 (the actual number of errors averaged) or 50000 (an approximation to the equivalent number of independent errors used). Thus, the Monte Carlo experiment of the phase-locked loop is more than good enough for a reliable bench-mark performance.

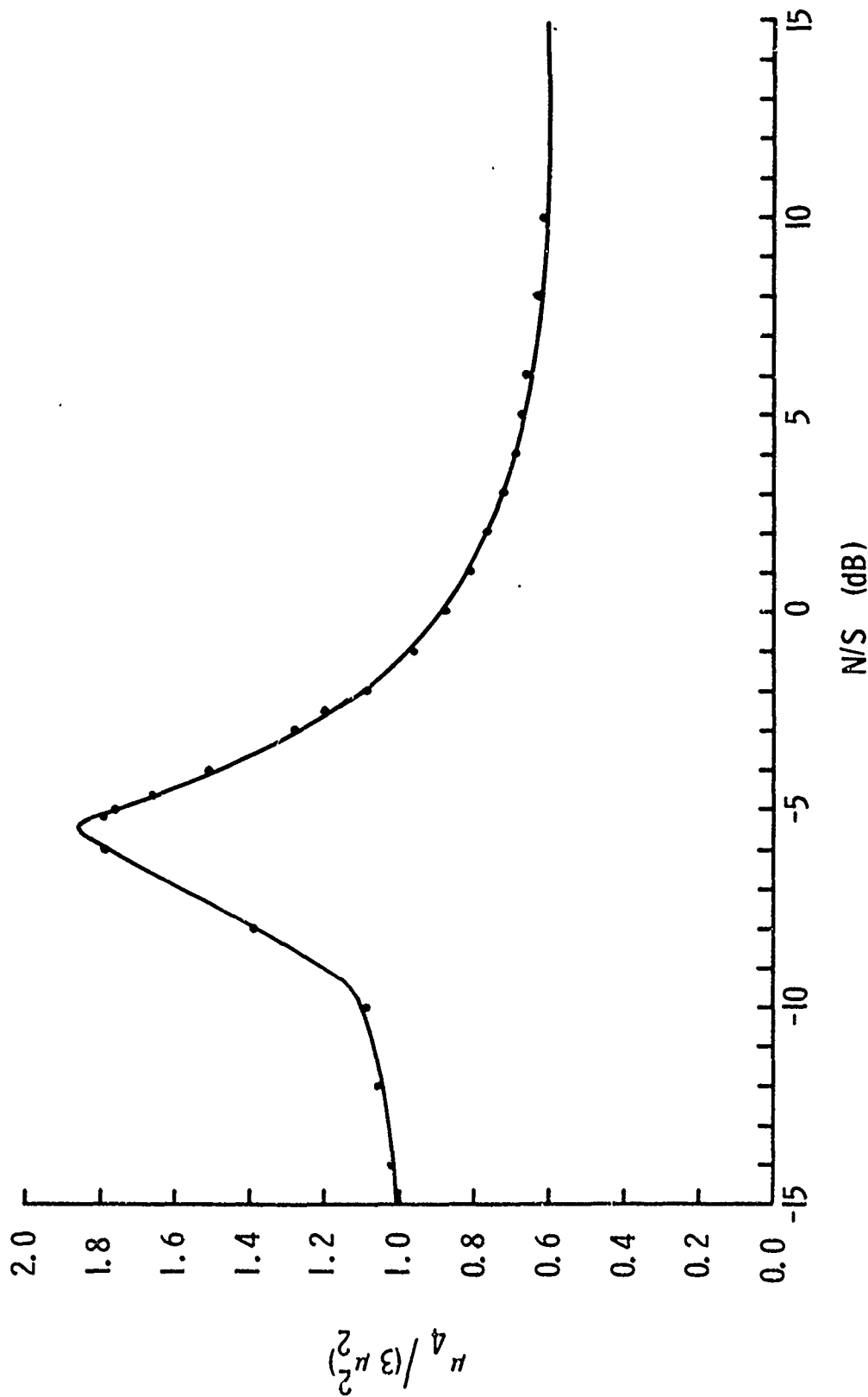


Fig. A-2. Fourth Moment Divided by Three Times the Squared Variance  
for the Phase-Locked Loop Error

Table A-1 Confidence Intervals for Linearized Filter

N/S (dB)	$\mu_2$	$\mu_4$	$\rho = \frac{\mu_4}{2\mu_2^2}$	Confidence (dB)	
				N=50000	N=500000
-14.0	.0372	.00424	3.06	$\pm 0.0828$	$\pm 0.0264$
-12.0	.0599	.0114	3.16	$\pm 0.0848$	$\pm 0.0270$
-10.0	.0993	.0321	3.26	$\pm 0.0867$	$\pm 0.0276$
-8.00	.173	.125	4.18	$\pm 0.103$	$\pm 0.0327$
-6.00	.343	.632	5.37	$\pm 0.120$	$\pm 0.0383$
-5.19	.480	1.24	5.38	$\pm 0.120$	$\pm 0.0384$
-5.00	.520	1.43	5.29	$\pm 0.119$	$\pm 0.0380$
-4.59	.606	1.83	4.98	$\pm 0.115$	$\pm 0.0366$
-4.01	.747	2.53	4.53	$\pm 0.108$	$\pm 0.0345$
-3.01	1.05	4.23	3.84	$\pm 0.0971$	$\pm 0.0309$
-2.50	1.17	4.95	3.62	$\pm 0.0933$	$\pm 0.0297$
-2.00	1.36	6.08	3.29	$\pm 0.0873$	$\pm 0.0278$
-1.94	1.39	6.26	3.24	$\pm 0.0863$	$\pm 0.0275$
-1.00	1.67	8.02	2.88	$\pm 0.0792$	$\pm 0.0252$
0.00	1.92	9.67	2.62	$\pm 0.0735$	$\pm 0.0234$
1.00	2.14	11.1	2.42	$\pm 0.0689$	$\pm 0.0219$
2.00	2.36	12.6	2.26	$\pm 0.0649$	$\pm 0.0206$
3.00	2.53	13.9	2.17	$\pm 0.0626$	$\pm 0.0199$
4.00	2.67	14.8	2.08	$\pm 0.0601$	$\pm 0.0191$
5.00	2.77	15.6	2.03	$\pm 0.0587$	$\pm 0.0187$
6.00	2.87	16.4	1.99	$\pm 0.0576$	$\pm 0.0183$
8.00	3.04	17.6	1.90	$\pm 0.0549$	$\pm 0.0174$
10.00	3.12	18.1	1.86	$\pm 0.0537$	$\pm 0.0171$

## Appendix B. Numerical Experiments With The Hermite Polynomial Expansion

The objective of the numerical methods investigated in this research were to:

- 1) devise methods to realistically (small computation time per estimate) solve the nonlinear filter problem using contemporary computers, and
- 2) to demonstrate that a practical nonlinear filter can be constructed to take advantage of its inherent accuracy as compared to a linear filter. Both objectives were achieved using the methods of the previous sections and demonstrated with the programs given (Hecht [6]).

The initial numerical data given is that generated at the University of Southern California, using an IBM 360-65 computer, and reported in [6]. The subsequent data was generated at Kirtland AFB, New Mexico, using a CDC 6600 computer and is reported here for the first time.

The first goal, reduced computation time, was demonstrated by comparison to Bucy and Senne [5], where a two-dimension problem, roughly equivalent to the phase-lock problem, was solved. In the Bucy and Senne paper sophisticated techniques were used to reduce the number of computations per estimate, reducing the computation time by a factor of 200. The reduced number of computations, per estimate, was approximately  $13 \times 10^3$ . For the filter using Hermite expansions, described in this paper, the number of computations per estimate was  $18 \times 10^2$ . However, the critical calculation, the evaluation of an exponential function, had to be done only 100 times per estimate. It would be indicated, therefore, that a time improvement of a factor of 130 could be expected. The Bucy and Senne paper gave results using a Burroughs B5500 computer, and had a measured time per estimate of 45 seconds. The Hermite

numerical results, given in this chapter, were obtained using an IBM 360 Model 65 computer. For the Monte Carlo experiments described in this Appendix there were 203 sample functions, each consisting of 130 points, which ran in 120 minutes, or approximately 0.273 seconds per estimate. Assuming the Burroughs B5500 was approximately equivalent to the IBM 360-65, there was a measured improvement of 165 times. The measured time on the IBM 360-65 was just about what was predicted in the referenced paper for a parallel processing computer, if such a computer should become available at a future date.

The second goal, simulation and demonstration of accuracy, was accomplished by comparing the nonlinear filter, described in Section C, with the relinearized filter described in Section B. The Monte Carlo results obtained were possible only because of the efficiency of the Hermite method. In the following descriptive material the filter of Section B is referred to as the "linear" filter, and that of Section C as the "nonlinear" filter. Both the linear and the nonlinear filters were designed to estimate the phase angle for the phase coherent communications problem.

As indicated in the discussion of Section B, the estimate of the phase angle is required modulo  $2\pi$ . The construction of both filters was such as to attempt to track the absolute phase angle. When evaluating the filters, however, the error, modulo  $2\pi$ , was the value used.

The parameter,

$$p_{11} = E[(x_1 - \hat{x}_1)^2(t) | Z_Y, \gamma = -\infty, t]$$

was the independent parameter for all comparisons. The variances of the message-model and the observation noise were related to  $p_{11}$  as

was shown in Section B. Selections of the numerical values for the initial experiments is given in Table B-1. Viterbi [5] shows the parameter  $p_{11}(0)$  is also the inverse of the effective signal to noise ratio,  $N/S$ . Thus,  $p_{11}(0)$  had the two physical interpretations:

- 1) equilibrium error variance.
- 2) effective noise to signal ratio.

A preliminary test was made to verify that both the linear and nonlinear filters were working properly. For small  $p_{11}$  one would expect both filters to give equal results. A sample function of 130 points (13 filter time constants) was generated for  $p_{11}^{1/2} = .1, .01$ , and  $.004$ ; the computer listings were given in Hecht [3]. The following results were noted:

- 1) The sequence of estimates for the linear and nonlinear filter agree with each other to about

$$\begin{aligned} 10^{-2} \text{ radian for } p_{11}^{1/2} &= .1 \\ 10^{-4} \text{ radian for } p_{11}^{1/2} &= .01 \\ 10^{-5} \text{ radian for } p_{11}^{1/2} &= .004 \end{aligned}$$

- 2) The measured variances and errors agree to within the same precision as the estimates.
- 3) The equilibrium computed and measured variances for the two filters agree with each other and with the linear computed value using the equations of Section B.

The nonlinear filter was tested at  $\bar{p}_{11}^{1/2} = .55$  ( $\bar{p}_{11} = .3025$ ,  $N/S = -5.2$  db) where the difference between the measured and theoretical linear variance was 3.5 db. The nonlinear filter simulation test at

Table B-1

## Numerical Values Used for Computer Simulations

$\Delta$	=	time between samples
$F$	=	filter time constant
$\frac{\Delta}{F}$	=	0.1
$\bar{P}_{11}(0)$	=	equilibrium continuous linear position error variance
$\bar{P}_{22}(0)$	=	equilibrium continuous linear velocity error variance
$E(x_1(0))$	=	0
$E(x_2(0))$	=	0
$E(x_1^2(0))$	=	$4\bar{P}_{11}(0)$
$E(x_2^2(0))$	=	$4\bar{P}_{22}(0)$
$q$	=	continuous message driving variance = 0.01
Number of points in each dimension for Gauss-Hermite numerical integration = 10		
Highest order of terms in series expansion of density function = 5		

$N/S = 5.2$  was under the same conditions as the linear filter, i.e., same noise sequences, but only 200 sample functions. The phase error variance for the nonlinear filter at these conditions was  $-3.55$  db, or  $1.45$  db better than the linear filter. This point is also shown in Figure B-1.

The cumulative average variance was plotted for both filters to show the stabilization of the average as a function of number of sample functions, Figure B-2. Points are plotted for every fifth sample function up until sample function number 100, then one final point at sample function number 200.

Large errors for both the linear and nonlinear filters for the phase angle problem are due to the phenomena of "cycle slippage," which occurs more frequently as  $N/S$  gets larger. The improvement in performance of the nonlinear filter was due, primarily, to the ability of the nonlinear filter to reduce the number of cycle slips. Sample function number 6 was identified as one in which the linear filter slipped several cycles whereas the nonlinear filter held on. The sequence of estimates and errors for both filters for this sample function were analyzed and a portion of sample function number 6 (from about  $N = 15$  to 52) is shown in Figure B-3 and the absolute value of the error, modulo  $2\pi$ , is plotted in Figure B-4. The figures show the linear filter has slipped a cycle at  $N=35$ , and is slipping a second cycle at  $N=50$ . The nonlinear filter, at the same time, has a large error at  $N=35$  but appears to recover nicely and by  $N=50$  it is tracking very well.

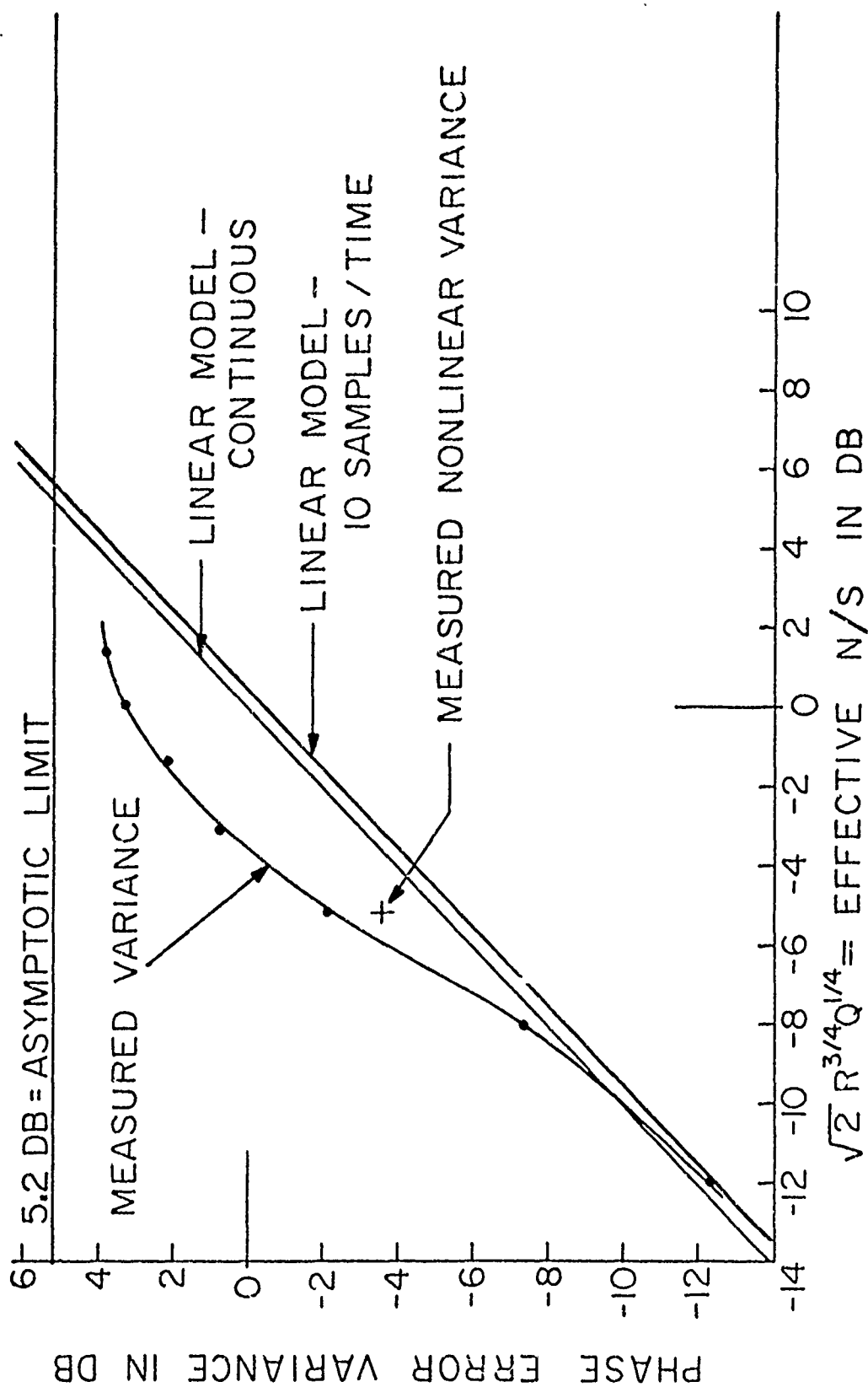


Fig. B-1. Hermite Expansion Error Summary

$$P(o) = 4P(\infty)$$

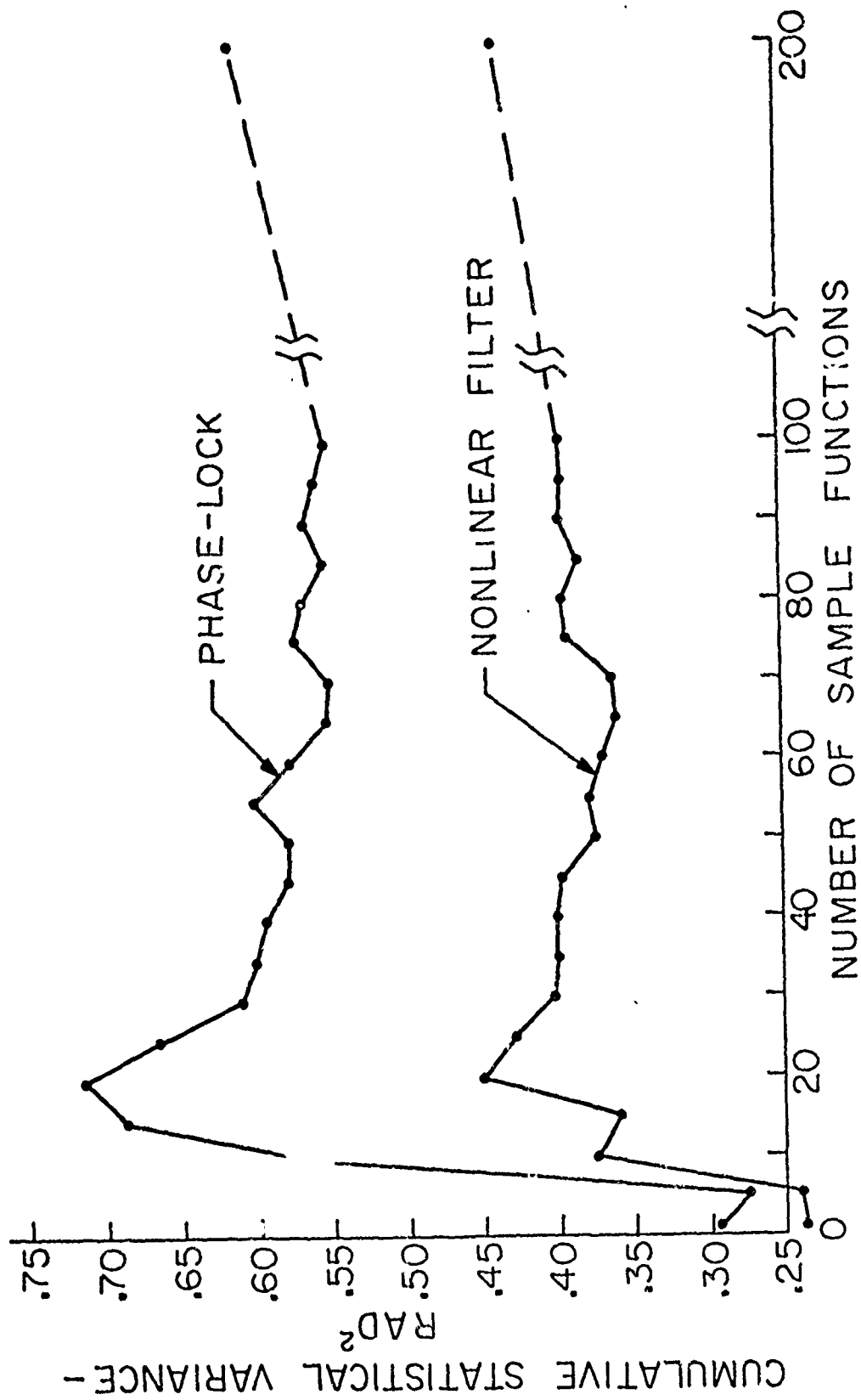


Fig. B-2. Cumulative Statistical Variance

$$P(\omega) = 4P(\omega)$$

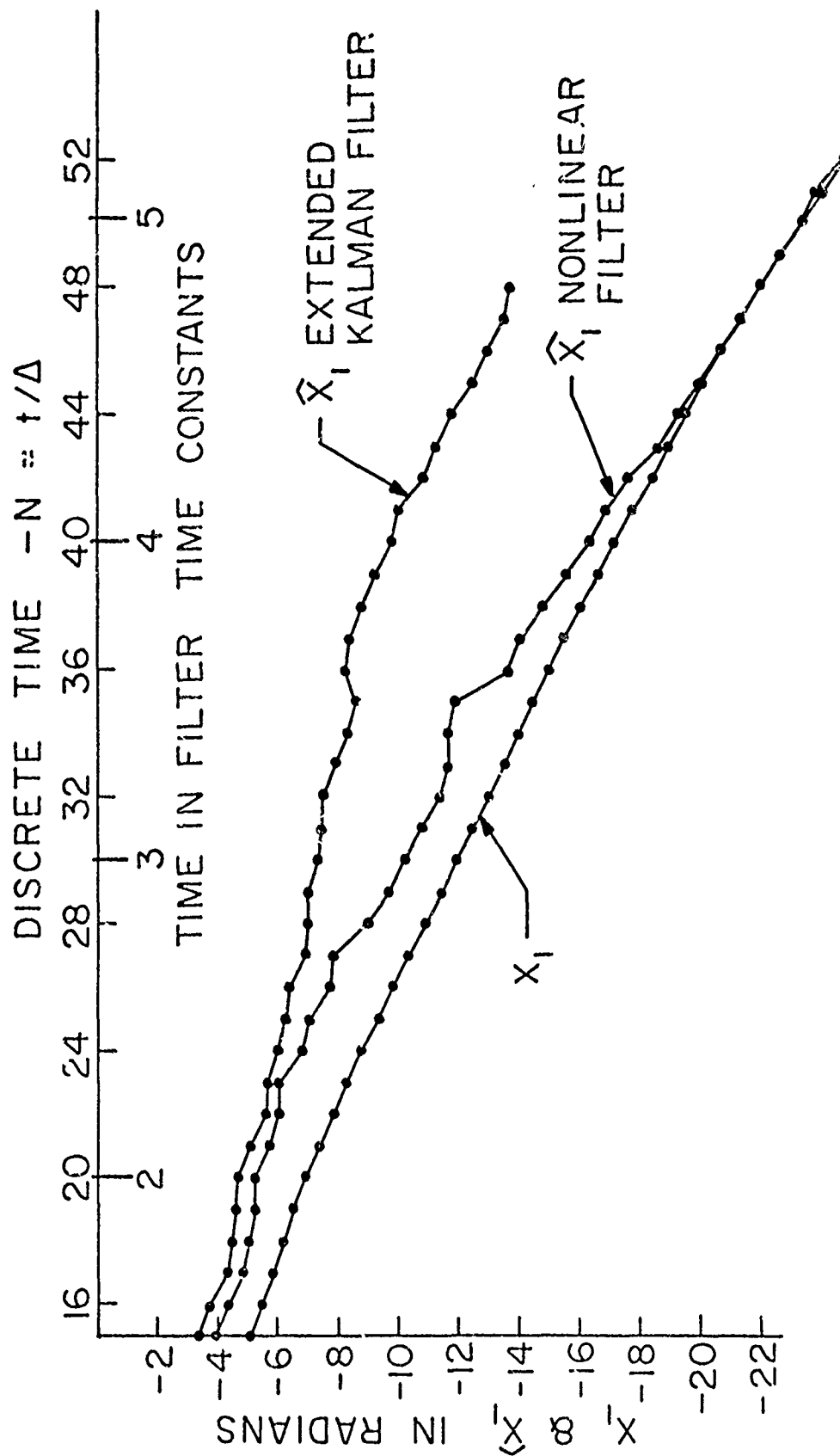


Fig. B-3. Portion of Sample Function No. 6

$$P_{11}(0) = 0.3025$$

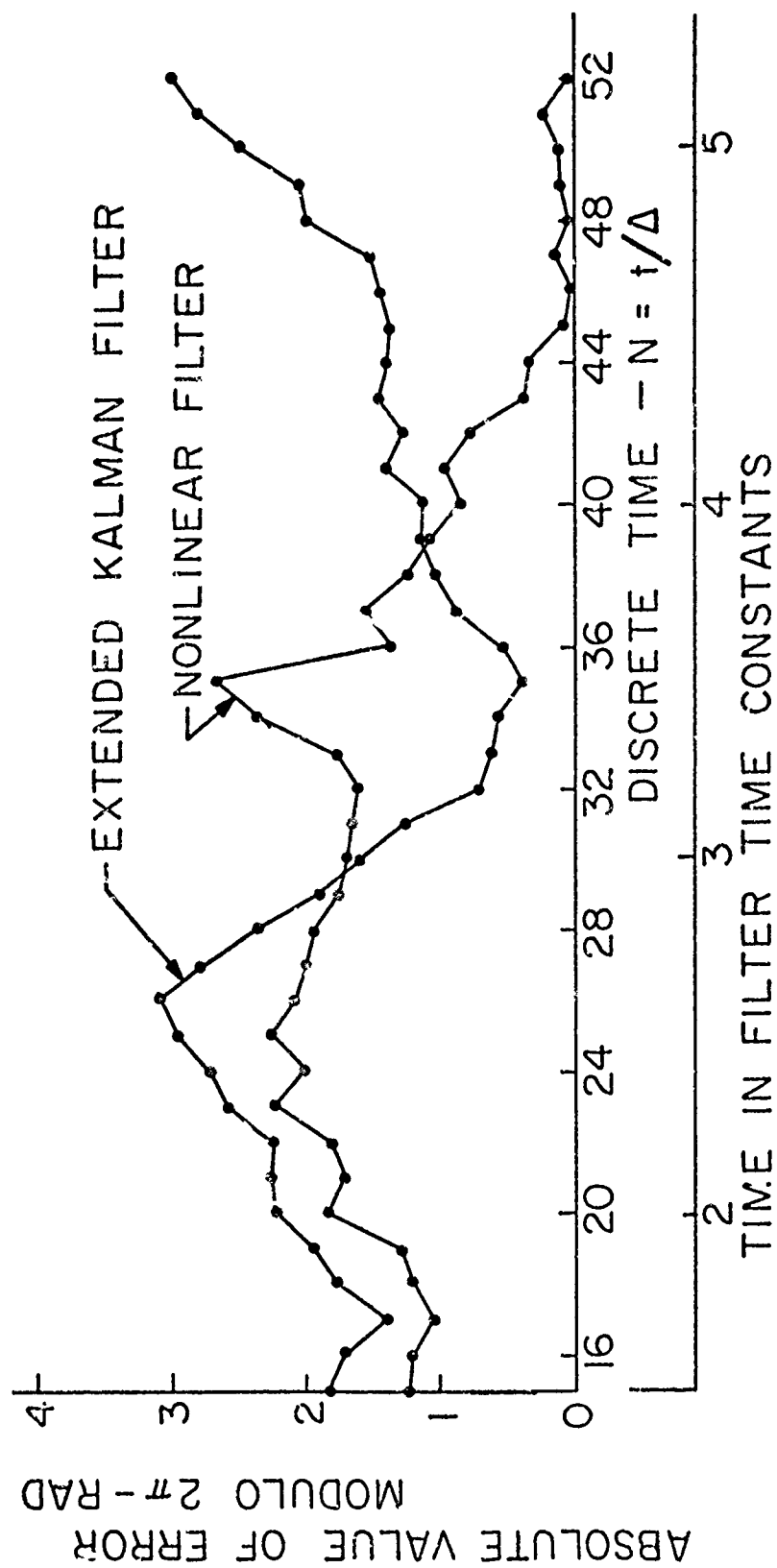


Fig. B-4. Error for Sample Function No. 6

$$P_{11}(0) = 0.3025$$

It can be seen that the measured variance in Figure B-1 is significantly different than the corresponding curve given in Appendix A (Figure A-1). Investigations showed that the initial conditions affected the variance for substantially longer than 3 time constants, as was originally assumed. The variances given in Figure B-1 were based on the variance of the initial estimate being four times the equilibrium variance, whereas Figure A-1 was based on the initial variance set equal to the equilibrium variance.

To demonstrate that the equilibrium solution could be achieved independent of the starting condition, one long sequence was run (83,000 points) starting at four times the equilibrium value. This is shown in Figure B-5, where markers are inserted to show the result for the two curves previously mentioned for the conditions given on the graph. After about 8000 time constants (80,000 points) the cumulative average appears to be approaching the solution given in Figure A-1, which was based on 5000 Monte Carlo sample functions of 130 points each, starting at the equilibrium value.

With this new knowledge, the Hermite nonlinear filter was again evaluated with 400 Monte Carlo functions with the same parameters except that the starting values were the computed equilibrium values. This new sequence was compared to the phase-lock results for an identical set of sequences. The cumulative errors for the two filters are plotted in Figure B-6, in a manner similar to Figure B-2. Both the phase-lock and the nonlinear filter had smaller errors than before. The phase-lock for 400 functions was -3.10 db and the nonlinear filter was -4.06 db, for an improvement of 0.96 db. Because of the larger number of samples the  $3\sigma$  confidence probability improved to about

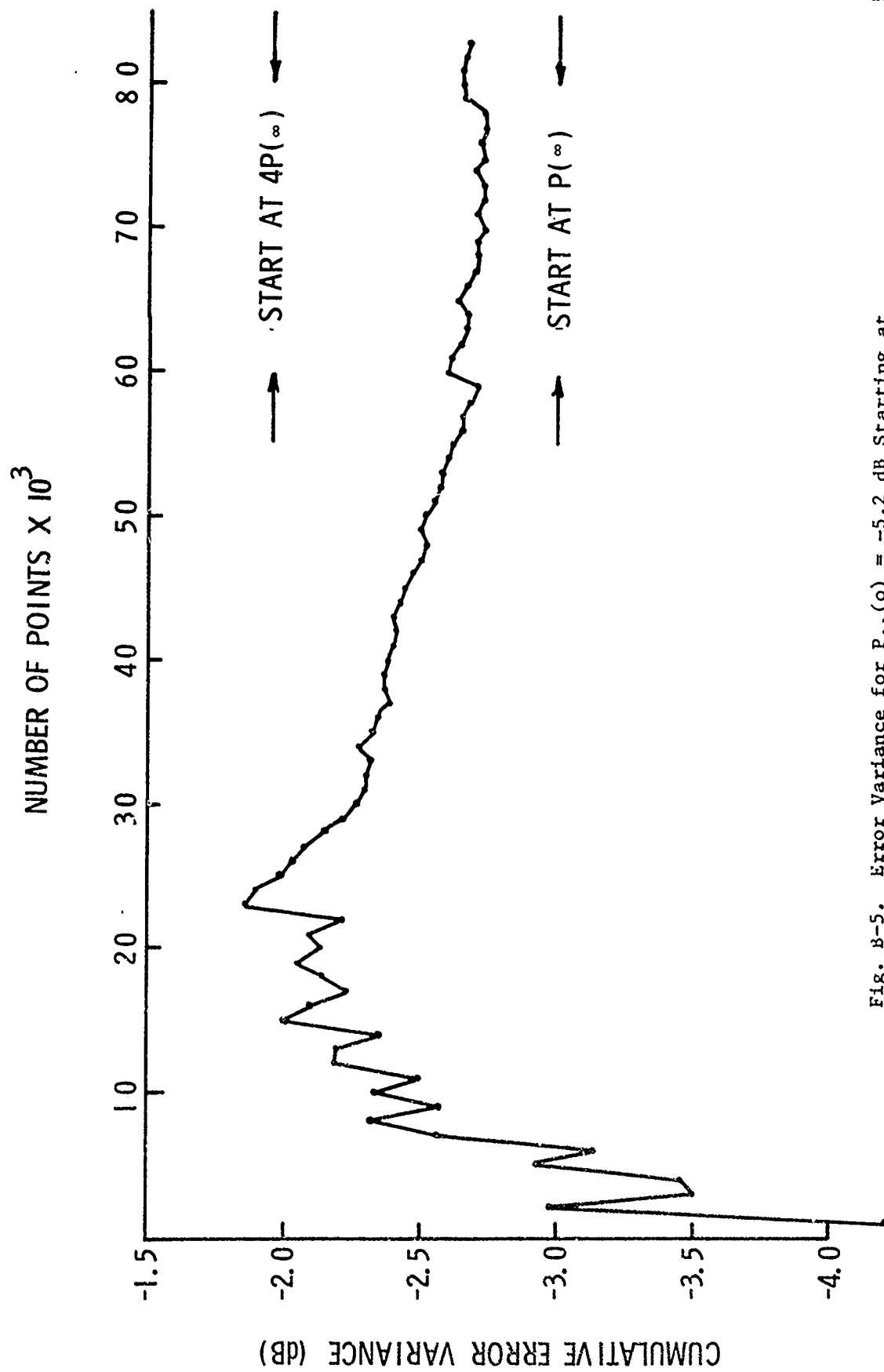


Fig. B-5. Error Variance for  $P_{11}(o) = -5.2$  dB Starting at  
 $P_{11}(o) = 4 P_{11}(\infty)$

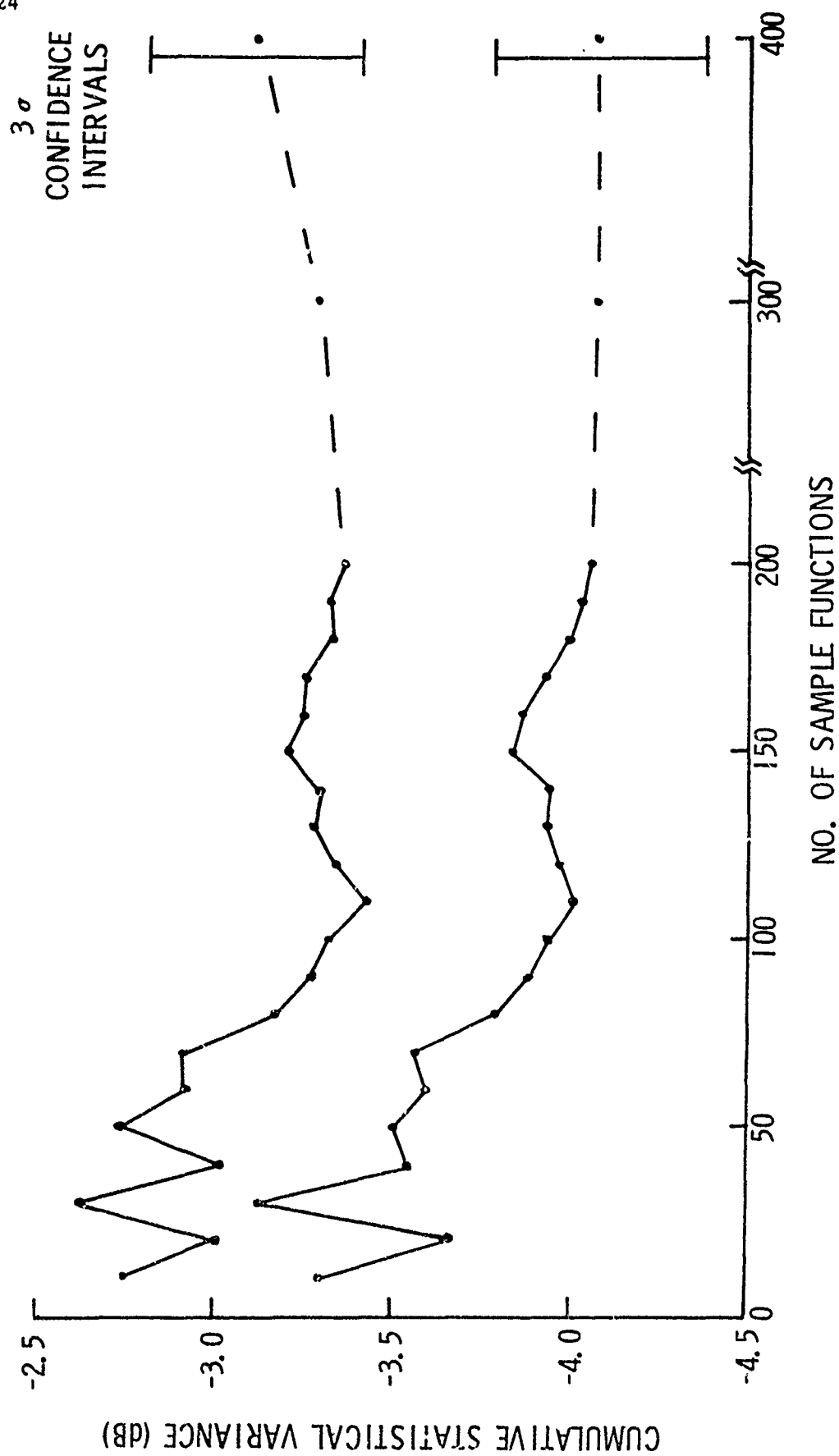


Fig. B-6. Cumulative Statistical Variance for  $P_{11}(o) = -5.2$  dB

$\pm 0.07$  for both filters. It was noted that the phase-lock error for 400 functions was 0.49 (-3.10 db) as compared to 0.50 (-3.0 db) for 5000 functions, or within .02 ( $.01/.49 = .0204$ ) of what might be considered the "correct" error.

The above data was generated on the CDC 6600 computer, and it was noted the nonlinear filter (Hermite expansion) computed the estimates at a rate of .127 Sec/estimate, which was more than two times faster than the previous computer.

In conclusion we note that a practical two-dimensional nonlinear filter was simulated using a digital computer. The digital filter error variance was within approximately 10% of the continuous model. Using Gauss-Hermite integration and Hermite Series expansions the nonlinear filter computed solutions to a phase angle problem at the measured rate of 0.273 seconds per estimate on a medium speed contemporary computer and .127 seconds per estimate on a high speed computer, which was 165 times faster than a roughly equivalent problem using the most advanced digital techniques prior to this paper. The phase angle problem solved was a model of an existing type of communications receiver which presently uses linearization methods to handle the nonlinearities. The simulated nonlinear filter using Hermite expansions showed an error variance reduction of .96 db at moderately high noise to signal ratio, with greater reductions at higher noise levels shown for other nonlinear filter methods. (See the discussion in the following Appendices.)

### Appendix C. Cyclic Point-Mass Experiments

In this appendix we will discuss a point-mass realization of the cyclic phase density recursion. We will show how the cyclic representation of the problem is substantially better behaved for high-noise applications than the Hermite expansion, which suffers from multiple modes. We begin with a description of the special case of the cyclic point-mass filter. The density recursion satisfies the relation (see the main chapter).

$$\tilde{J}_n \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = S(y_1) \int_{-\frac{\pi}{\Delta}}^{\frac{\pi}{\Delta}} \sum_{v_2=-\infty}^{\infty} \exp \left\{ -\frac{1}{2q\Delta} \left( y_2 - \eta + \frac{2\pi}{\Delta} v_2 \right)^2 \right\} \tilde{J}_{n-1} \begin{pmatrix} y_1 - \eta\Delta \\ \eta \end{pmatrix} d\eta \quad (C-1)$$

$$\begin{aligned} \text{where} \quad S(y_1) &\triangleq C_0 \exp \left\{ -\frac{\Delta}{2r} \left[ (z_1 - \cos y_1)^2 + (z_2 - \sin y_1)^2 \right] \right\} \\ &= C_1 \exp \left\{ +\frac{\Delta}{r} (z_1 \cos y_1 + z_2 \sin y_1) \right\}, \end{aligned}$$

and

$$-\pi \leq y_1 < \pi, \quad -\frac{\pi}{\Delta} \leq y_2 < \frac{\pi}{\Delta}.$$

The estimate which was simulated for the cyclic densities was the cyclic estimate described in the main chapter.

For large values of  $|v_2|$  in the summation

$$F(y_2, \eta) = \sum_{v_2=-\infty}^{\infty} \exp \left\{ -\frac{1}{2q\Delta} \left( y_2 - \eta + \frac{2\pi}{\Delta} v_2 \right)^2 \right\}$$

the expression is negligibly small.

Preceding page blank

Therefore, only those integers  $v_2$  were used where

$$\max_{y_2, \eta} F(y_2, \eta) > \sim 10^{-20}.$$

The program that was developed makes the above test on  $F(y_2, \eta)$ , and in all of the results to date only the value  $v_2 = 0$  has been found to be significant.

$F(y_2, \eta)$  is not a function of  $n$ , the integer time, and therefore was computed only one time in advance and stored for use in (C-1) for all  $n$ . A further simplification was made by taking advantage of the fact that  $F(y_2, \eta)$  is only a function of  $y_2 - \eta$ . That is, after discretizing the argument  $y_2$ ,  $F(y_2, \eta)$  is computed for a range of values of  $y_2 - \eta$ , rather than for all combinations of  $y_2$  and  $\eta$ . In the sequel we let the running variable  $\eta$  be called  $x_2^*$ .

The 2-dimensional interval

$$-\pi \leq x_1 < \pi,$$

and

$$-\frac{\pi}{\Delta} \leq x_2 < \frac{\pi}{\Delta}$$

is divided into  $m$  and  $n$  equally spaced sub-intervals, respectively, and each sub-interval is defined by a point on its center,  $y_{1_i}, y_{2_j}$ , with  $i = 1, \dots, m$  and  $j = 1, \dots, n$ . The points  $y_{1_i}$  and  $y_{2_j}$  define a grid which remains constant with respect to time, and the density function is represented by point masses defined only on this grid, where the magnitude of the point masses approximates the density at that point. From (C-1), the points  $x_{2_j}^* = y_{2_j}$  ( $j = 1, n$ ); that is, for each integer  $j$ , the two grid points are identical.

The integrand in (C-1) needs to be evaluated only for those  $x_{2j}^*$  where  $\max_{y_{2i} - x_{2j}^*} F(y_{2i}, x_{2j}^*) > \text{approximately } 10^{-20}$ . For computing

an updated density function  $\tilde{J}_n \left( \begin{smallmatrix} y_{1i} \\ y_{2j} \end{smallmatrix} \right)$  the integrand needs to be

evaluated only a small number of times for each  $y_{2j}$  (typically 10 times for  $n = 200$ ).

The main difficulty in the mechanization of (C-1) is in determining  $\tilde{J}_{n-1} \left( \begin{smallmatrix} y_{1i} - x_{2j}^* \Delta \\ x_{2i}^* \end{smallmatrix} \right)$ , having the prior value of this function available only

at a discrete set of in points in the first argument, different than those defined by  $y_{1i} - x_{2j}^* \Delta$ .

One approach to solving this problem is as follows. Let

$$\begin{aligned} x_{1i} &= -\pi + 2\pi \frac{(i-1)}{m} + \frac{1}{2} \left( \frac{2\pi}{m} \right) & i = 1, m \\ x_{2j}^* &= -\frac{\pi}{\Delta} + \frac{2\pi}{\Delta} \frac{(j-1)}{n} + \frac{1}{2} \left( \frac{2\pi/\Delta}{n} \right) & j = 1, n \end{aligned} \quad (C-2)$$

as defined above.

Now,

$$\begin{aligned} y_{1i} - x_{2j}^* \Delta &= -\pi + 2\pi \frac{(i-1)}{m} + \frac{1}{2} \left( \frac{2\pi}{m} \right) - \Delta \left[ -\frac{\pi}{\Delta} + \frac{2\pi}{\Delta} \frac{(j-1)}{n} + \frac{1}{2} \left( \frac{2\pi/\Delta}{n} \right) \right] \\ &= 2\pi \left( \frac{i-1}{m} - \frac{j-1}{n} \right) + \frac{1}{2} \left( \frac{2\pi}{m} - \frac{2\pi}{n} \right) \\ &= 2\pi \left( \frac{i-1}{m} - \frac{j-1}{n} \right) + \pi \left( \frac{1}{m} - \frac{1}{n} \right) \end{aligned} \quad (C-3)$$

We want to use the integers  $i$  and  $j$  to define a new grid point  $k$ , on the  $x_1$  axis; the  $k$  grid point must agree with an original  $x_{1i}$

grid point. That is, from (8) and (9) we want

$$x_{1k} = y_{1i} - x_{2j}^* \Delta$$

or

$$-\pi + 2\pi \left( \frac{k-1}{m} \right) - \frac{1}{2} \left( \frac{2\pi}{m} \right) = 2\pi \left[ \left( \frac{i-1}{m} \right) - \left( \frac{j-1}{n} \right) \right] + \pi \left( \frac{1}{m} - \frac{1}{n} \right) \quad (C-4)$$

from which

$$k = i - \frac{m}{n} j + \frac{1}{2} \left( \frac{m}{n} + m \right) \quad (C-5)$$

In the initial approach to this problem, when  $k$  as computed from (C-5) was not an integer, the nearest integer value was selected. In order to assure an  $x_1$  grid point falling exactly on  $x_1 = 0$  m must be an odd integer. It is also desirable to have  $k$  be correct (an integer value from (C-5)) when the  $x_2^*$  grid point  $j$  is in the center of its range, requiring  $n$  to be an odd number. To meet the above requirements, and to subdivide such that other points,  $k$ , might match exactly with some  $i$ , the ratio  $\frac{n}{m}$  should be an odd integer, which also assures that no  $k$  point will fall exactly in the middle of two adjacent  $i$  points.

A modification was made to the formula for computing  $k$ , to account for the possibility of  $k$ , as determined from (C-5), not falling in the range of  $(1, m)$ .

For

$$\begin{aligned} k > m & \quad k^* = k - m \\ k < 1 & \quad k^* = k + m \end{aligned} \quad (C-6)$$

Equation (C-6) is equivalent to folding back on the  $x_1$  axis to remain within the interval  $[-\pi, \pi)$ .

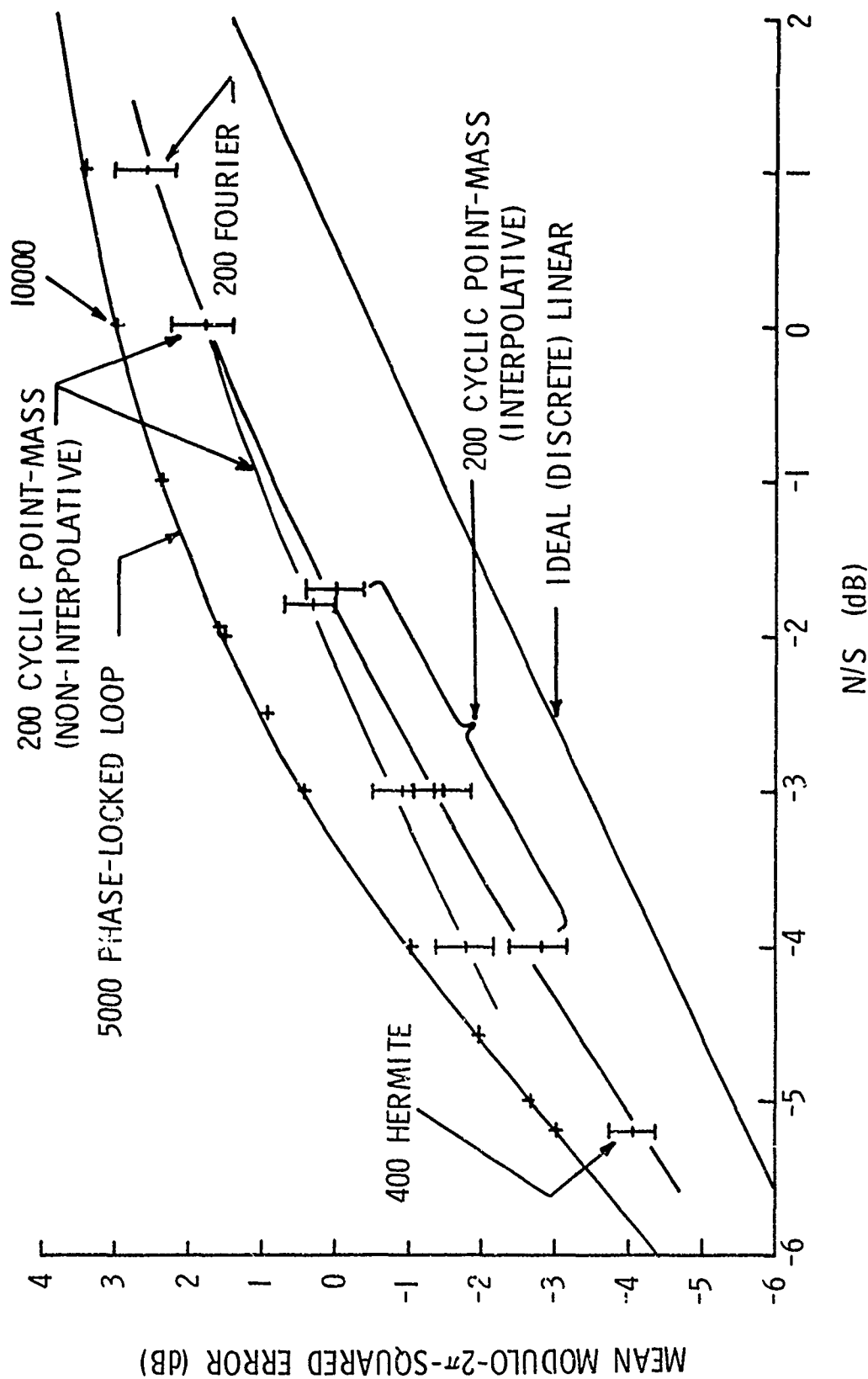


Fig. C-1. Nonlinear Filter Summary (Enlarged)

The above technique works satisfactorily but requires a large number of grid points to stabilize on the "true" nonlinear estimate, necessitating substantial computer expense. To test for convergence, a fixed noise sequence was generated and the filter was used to estimate the state with the number of grid points progressively increased until increasing the number of points caused no change in the sequence of estimates. It was subsequently learned that the number of points needed for convergence varied as a function of the  $N/S$  ratio. Substantial data was generated with the above method. Fig. C-1 shows the phase error variance as the upper of the two curves, based on Monte Carlos of 200 independent sample paths of 100 steady state points. The 3 $\sigma$  confidence intervals represent 2000 points (one each time constant).

An improvement which significantly reduced the computational requirements was to let  $k$ , from (11), take on non-integer values. The value of the density function at these places was evaluated by linear interpolation between the two adjacent integer values of  $k$ , where the density function was available from the prior cycle. The interpolation was required only in the  $x_1$  direction, since the densities in the  $x_2$  coordinate direction are computed at exactly the places where they are needed for the recursion formula. For interpolating between  $k = m$  and  $k = l$ , the points  $m$  and  $l$  were considered adjacent, completing a circle. The convergence test described above was applied to the modified filter and stabilized for a substantially smaller number of grid points. The lower curve of Fig. C-1 was generated using the filter with interpolation.

The two curves show the same filter error variance at about  $N/S = 0$  dB, which was in the vicinity of the  $N/S$  ratio where the convergence tests were made prior to introducing the interpolation. For lower  $N/S$  the interpolating filter shows lower errors, being about one dB. less at about  $N/S = -4$  dB. The numerical granularity associated with non-interpolation apparently causes significant errors at the lower  $N/S$  ratios, due to the sharpness of the phase error density function. At higher  $N/S$  ratios, the filter error density is so diffuse that the numerical errors are of no consequence.

The numerical results of the Monte Carlo experiments for the cyclic point-mass filter are given in Table C-1. Table C-2 gives the associated improvement of the cyclic filter over the phase-locked loop results reported in Appendix A, and Table C-3 gives the difference between the nonlinear filters and the ideal linear analysis. In all cases the minus and plus  $3\sigma$  confidence intervals are given, where for 2000 points the lower threshold is -0.394 dB below nominal and the upper threshold is 0.433 dB above the nominal.

Table C-1. Monte Carlo Mod  $2\pi$  Error Performance Data for the Cyclic Point Mass Estimates

N/S (dB)	Non-Interpolative MSE (dB) (Mod $2\pi$ )			Interpolative MSE (dB) (Mod $2\pi$ )		
	Low	Nom.	High	Low	Nom.	High
-4.01	-2.20	-1.81	-1.38	-3.22	-2.83	-2.40
-3.01	-1.33	-0.94	-0.51	-1.87	-1.48	-1.05
-1.79	-0.09	0.30	0.73	----	----	----
-1.70	----	----	----	-0.38	0.01	0.44
0.00	1.46	1.85	2.28	----	----	----

Table C-2. Monte Carlo Improvements  
Cyclic Point-Mass over Phase-Locked Loop

N/S (dB)	Non-Interpolative			Interpolative		
	MSE (dB) Improvement			MSE (dB) Improvement		
	Low	Nom.	High	Low	Nom.	High
-4.01	0.33	0.76	1.15	1.35	1.78	2.17
-3.01	0.93	1.36	1.75	1.47	1.90	2.29
*-1.79	1.02	1.45	1.84	----	----	----
** -1.70	----	----	----	1.41	1.84	2.23
0.00	0.75	1.18	1.57	----	----	----

\* Phase-Locked Performance read from graph (1.75 dB)

\*\* Phase-Locked Performance read from graph (1.85 dB)

Table C-3. Monte Carlo Difference  
Between Cyclic Point-Mass and Ideal Linear

N/S (dB)	Non-Interpolative			Interpolative		
	MSE (dB) Difference			MSE (dB) Difference		
-4.01	2.23	2.62	3.05	1.21	1.60	2.03
-3.01	2.10	2.49	2.92	1.56	1.95	2.38
-1.79	2.13	2.52	2.95	----	----	----
-1.70	----	----	----	1.75	2.14	2.57
0.00	1.9	2.29	2.72	----	----	----

The summary of these data in Fig. C-1 is shown with the phase-locked loop performance and the Idealized Linear reference curve. Fig. C-2

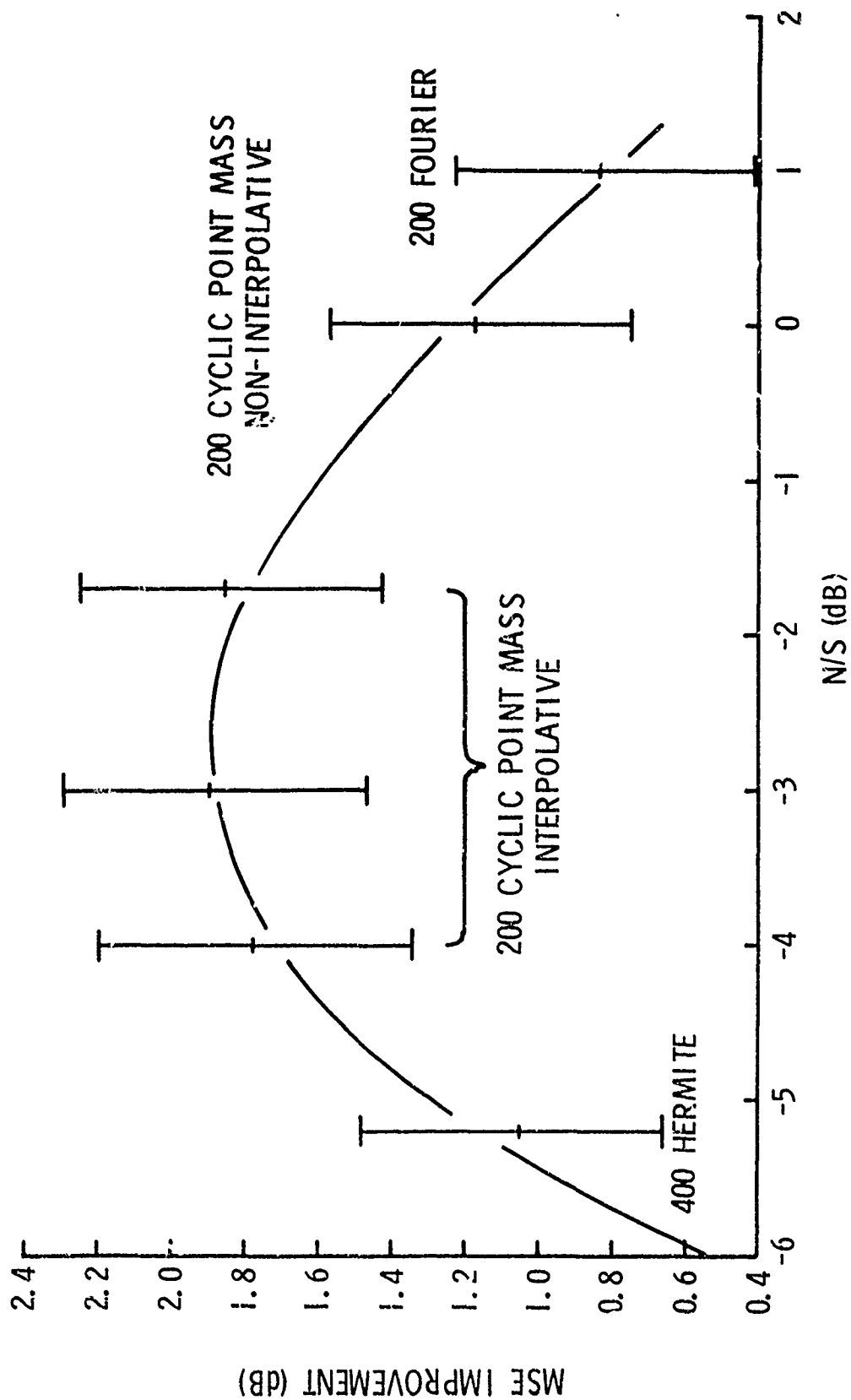


Fig. C-2. MSE Improvement of Nonlinear Filters over Phase-Locked Loop

shows the improvements over the phase-locked loop and Fig. C-3 shows the difference between nonlinear and ideal. In all figures the Hermite point from Appendix B is superimposed.

In addition to determining the Monte Carlo performance, a study was done to determine the execution times for various grid sizes in an effort to obtain a cost versus performance comparison.

The results of Fig. C-1 were obtained with a grid of  $m = 21$  and  $n = 105$ . The number  $p$  is evaluated by the program for each problem condition and represents the number of points on each side of the point  $(x_{1_i}, x_{2_j})$  in the  $x_2$  direction, which contribute to the computation of the density of that point.

An estimate of the time required to update the density function was based on the knowledge that the time was roughly proportional to the number of computations. For each  $m, n$  it requires  $m \times n \times p$  computations. The measured data for the  $21 \times 105$  grid case was about 0.215 seconds per estimate (as compared to the 5-coefficient Hermite value of 0.121 seconds per estimate). Using the above scaling, the data of Table C-4 was generated.

To determine an adequate grid size many runs were made with the same random sequence inputs, using different grid size combinations. All runs were made with  $N/S = -1.7$  dB, Filter Time constant = 5.13 seconds, samples per time constant = 10,  $\tau = .010$  and  $r = 1.735$ . The sequence of estimates for a ten-increment time period (integer time = 31, 40) were compared with each other. It was desired to find a combination which gave reasonable, good results (as compared to larger grids) while minimizing the time per estimate. Table C-5 gives 3 sequences with the grid size progressively increasing while maintaining the ratio

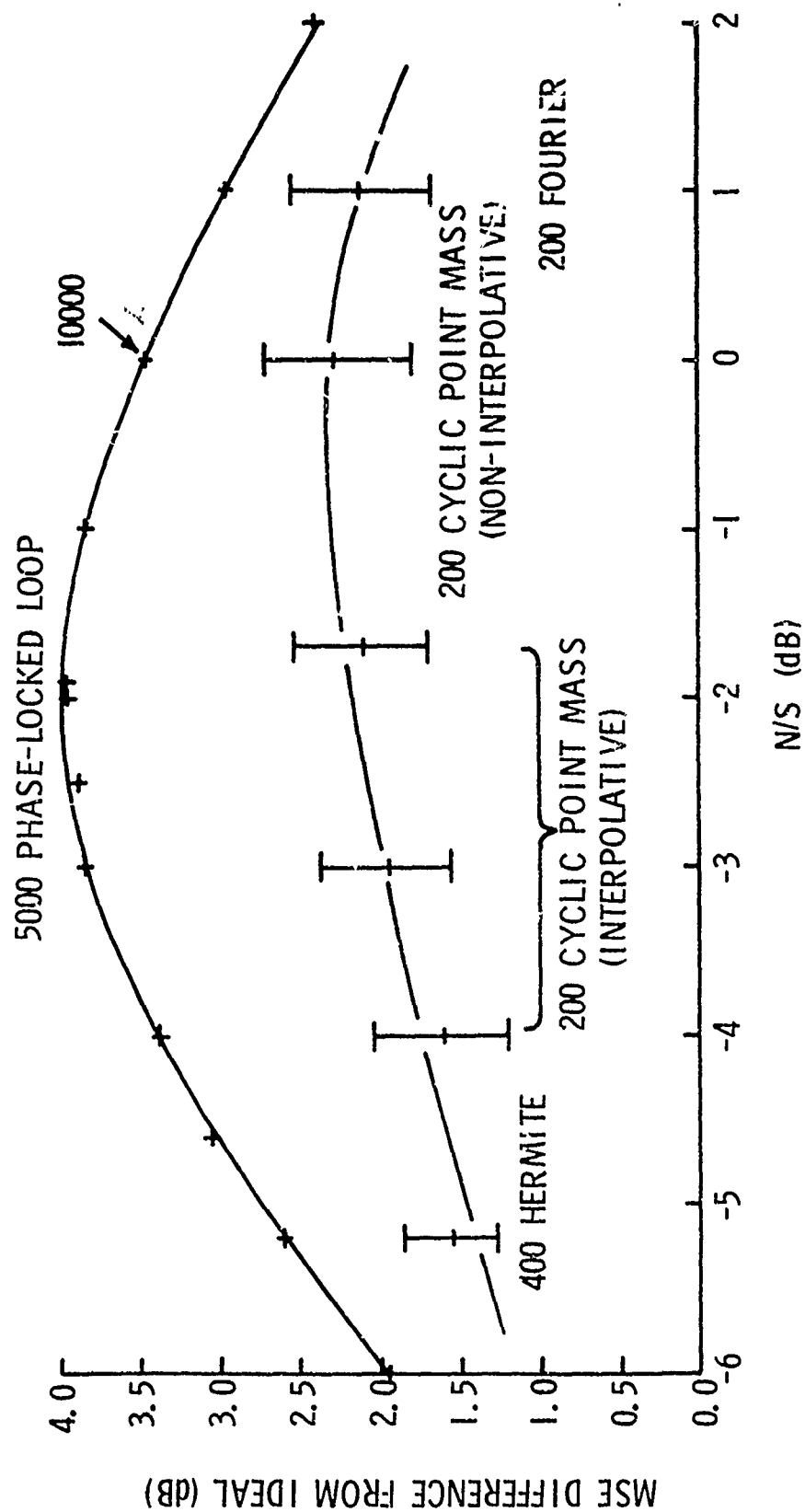


Fig. C-3. MSE Difference from Ideal Linear Analysis

$n/m = 5$ . Table C-6 maintains  $n = 155$  while varying the ratio  $n/m$  from 3 through 5. Table C-7 maintains  $m = 31$  while varying the ratio  $n/m$  from 5 through 7. Analyzing the absolute error sequences from Tables C-5 through C-7 we note that for  $n/m = 5$ , and increasing grid size, the errors become smaller as the grid size increases, with very little change between  $m = 31$  and  $n = 41$ , and insignificant differences between  $m = 21$  and  $n = 41$ . For  $n = 155$ , Table C-6, and for  $m = 31$ , Table C-7, the same general trend was observed. That is, slight but insignificant improvement with increasing grid size. For Table C-7, especially, there appears to be no advantage in increasing  $m$ . Tables C-5 and C-6 suggest, however, that ultimate stabilization can be achieved by increasing  $m$  with  $n$  about 155. The choice of a  $21 \times 105$  grid for the Monte Carlo experiments was based on a compromise between time and accuracy, as illustrated in the Tables.

Table C-4. Timing Estimates

Grid	p	m/n	Time, Est Sec.
21 x 105	4	5	.215
31 x 155	6	5	.700
41 x 205	8	5	1.63
31 x 93	4	3	.279
31 x 125	5	4	.468
31 x 155	6	5	.700
31 x 195	8	6	1.17
31 x 217	9	7	1.46
21 x 145	6	7	.442
25 x 151	6	6	.547
31 x 155	6	5	.700
39 x 155	6	4	.875
51 x 155	6	3	1.15

$m$  = number of lines in  $x_1$  direction

$n$  = number of lines in  $x_2$  direction

$2p + 1$  = number of computations to update each grid point

Table C-5.  $n/m$  Constant

$\frac{n}{m} = 5$		<u>.215 Sec.</u>		<u>.700 Sec.</u>		<u>1.63 Sec.</u>	
		$m = 21$		$m = 31$		$m = 41$	
		$n = 105$		$n = 155$		$n = 205$	
Time	$x_1$	$\hat{x}_1$	$ \hat{x}_1 - x_1 $	$\hat{x}_1$	$ \hat{x}_1 - x_1 $	$\hat{x}_1$	$ \hat{x}_1 - x_1 $
31	-2.394	3.094	.795	3.100	.789	3.106	.783
32	-2.296	-3.085	.789	-3.081	.785	-3.075	.779
33	-2.244	-2.146	.098	-2.216	.028	-2.234	.010
34	-2.220	-1.905	.315	-1.943	.277	-1.955	.265
35	-2.210	-2.146	.064	-2.163	.047	-2.172	.038
36	-2.221	-2.903	.682	-2.880	.659	-2.868	.647
37	-2.222	-1.908	.314	-1.912	.310	-1.921	.301
38	-2.198	-1.246	.951	-1.274	.924	-1.287	.911
39	-2.162	-1.312	.850	-1.330	.832	-1.343	.819
40	-2.195	-1.894	.301	-1.866	.329	-1.868	.327

Table C-6. n Constant

n = 155

Time	<u>.700 Sec.</u>		<u>.875 Sec.</u>		<u>1.15 Sec.</u>	
	m = 31		m = 39		m = 51	
	$\frac{n}{m} = 5$		$\frac{n}{m} \approx 4$		$\frac{n}{m} \approx 3$	
	$\hat{x}_1$	$ \hat{x}_1 - x_1 $	$\hat{x}_1$	$ \hat{x}_1 - x_1 $	$\hat{x}_1$	$ \hat{x}_1 - x_1 $
31	3.100	.789	3.105	.784	3.106	.783
32	-3.081	.785	-3.077	.781	-3.076	.780
33	-2.216	.028	-2.232	.012	-2.246	.002
34	-1.943	.277	-1.954	.266	-1.960	.260
35	-2.163	.047	-2.172	.038	-2.177	.033
36	-2.880	.659	-2.870	.649	-2.867	.646
37	-1.912	.310	-1.920	.302	-1.925	.297
38	-1.274	.924	-1.287	.911	-1.290	.908
39	-1.330	.832	-1.343	.829	-1.347	.825
40	-1.866	.329	-1.871	.324	-1.870	.323

Table C-7. m Constant

m = 31

Time	<u>.700 Sec.</u>		<u>1.17 Sec.</u>		<u>1.46 Sec.</u>	
	m = 155		m = 195		m = 217	
	$\frac{n}{m} = 5$		$\frac{n}{m} \approx 6$		$\frac{n}{m} \approx 7$	
	$\hat{x}_1$	$ \hat{x}_1 - x_1 $	$\hat{x}_1$	$ \hat{x}_1 - x_1 $	$\hat{x}_1$	$ \hat{x}_1 - x_1 $
31	3.100	.789	3.100	.789	3.101	.788
32	-3.081	.785	-3.081	.785	-3.081	.785
33	-2.216	.028	-2.216	.028	-2.212	.024
34	-1.943	.277	-1.943	.277	-1.942	.278
35	-2.163	.047	-2.163	.047	-2.162	.048
36	-2.880	.659	-2.879	.658	-2.878	.657
37	-1.912	.310	-1.912	.310	-1.912	.310
38	-1.274	.924	-1.274	.924	-1.275	.923
39	-1.330	.832	-1.330	.832	-1.331	.831
40	-1.866	.329	-1.867	.328	-1.869	.326

# Appendix D. A Fourier Series Experiment

Mallinckrodt, Bucy, and Cheng [7] have observed the fact that since the cyclic phase density is periodic, a Fourier Series appears appropriate for representation of the density functions. They have developed equations for the evolution of the Fourier Series for the one dimensional problem. We extend their analysis to our two-dimensional problem and present the preliminary results of a numerical experiment in this appendix.

We begin by observing that an arbitrary function  $\tilde{J}(y)$  periodic on the rectangle  $-\pi \leq x_1 < \pi$ ,  $-\frac{\pi}{\Delta} \leq x_2 < \frac{\pi}{\Delta}$  may be represented in terms of its two dimensional Fourier Series

$$\tilde{J}_n \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim \sum_m \sum_\ell a_{m\ell}^n e^{+im y_1} e^{+i\ell \Delta y_2}, \quad (D-1)$$

where

$$a_{m\ell}^n = \int_{-\frac{\pi}{\Delta}}^{\frac{\pi}{\Delta}} \int_{-\pi}^{\pi} \tilde{J}_n \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} e^{-imy_1} e^{-i\ell \Delta y_2} dy_1 dy_2 \quad (D-2)$$

Now the cyclic density obeys the recursion relation

$$\tilde{J}_n \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = S(y_1) \int_{-\frac{\pi}{\Delta}}^{\frac{\pi}{\Delta}} M(y_2 - x_2) \tilde{J}_{n-1} \begin{pmatrix} y_1 - x_2 \Delta \\ x_2 \end{pmatrix} dx_2, \quad (D-3)$$

where

$$S(y_1) = C_0 \exp \left\{ \frac{z_1 \cos y_1 + z_2 \sin y_1}{r/\Delta} \right\}, \quad (D-4)$$

Preceding page blank

and

$$M(u) = \sum_k \exp \left\{ - \frac{\left(u + \frac{2\pi k}{\Delta}\right)^2}{2q\Delta} \right\} \quad (D-5)$$

Expressing  $M(u)$  as a Fourier Series yields

$$M(u) = \sum_v m_v e^{iv\Delta u}, \quad (D-6)$$

where

$$m_v = \int_{-\frac{\pi}{\Delta}}^{\frac{\pi}{\Delta}} \sum_k \exp \left\{ - \frac{\left(u + \frac{2\pi k}{\Delta}\right)^2}{2q\Delta} \right\} e^{-iv\Delta u} du \quad (D-7)$$

$$\begin{aligned} &= \int_{-\infty}^{\infty} \exp \left\{ - \frac{u^2}{2q\Delta} + iv\Delta u \right\} du \\ &= 2\pi \left\{ \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp \left[ - \frac{u^2}{2q\Delta} \right] \exp \left[ -iu(-v\Delta) \right] du \right\} \\ &= 2\pi \left\{ \frac{\sqrt{q\Delta}}{\sqrt{2\pi}} \exp \left[ - \frac{q\Delta}{2} (-v\Delta)^2 \right] \right\} \\ &= \sqrt{2\pi q\Delta} \exp \left[ - \frac{qv^2\Delta^3}{2} \right] \end{aligned} \quad (D-8)$$

Next, we represent  $S(y)$  by an infinite series by making the substitution  $z_1 + iz_2 = |z| \exp(i\theta)$ , so that  $z_1 = |z| \cos \theta$ , and  $z_2 = |z| \sin \theta$ . Then  $S(y)$  is expressed as

$$\begin{aligned} S(y) &= C_0 \exp \left\{ \frac{|z|}{\Delta/r} \cos (y-\theta) \right\} \\ &= C_0 \sum_{\lambda} I_{\lambda} \left( \frac{|z|}{\Delta/r} \right) \exp \left\{ i\lambda (y-\theta) \right\}, \\ &= C_0 \sum_{\lambda} S_{\lambda} \exp \{ i\lambda y \} \end{aligned} \quad (D-9)$$

where  $I_\lambda$  is the modified Bessel function of imaginary argument of order  $\lambda$  (see Abramowitz and Stegun [1], Equation 9.6.34).

Finally, we combine the definition (D-2) with the expression (D-3), thereby obtaining

$$a_{m\ell}^n = \int_{-\frac{\pi}{\Delta}}^{\frac{\pi}{\Delta}} \int_{-\pi}^{\pi} S(y_1) \left[ \int_{-\frac{\pi}{\Delta}}^{\frac{\pi}{\Delta}} M(y_2 - x_2) \tilde{J}_{n-1} \left( \begin{matrix} y_1 - x_2 \Delta \\ x_2 \end{matrix} \right) dx_2 \right] e^{-imy_1} e^{-i\ell\Delta y_2} dy_1 dy_2$$

(let  $\tau = y_2 - x_2$ , or  $y_2 = \tau + x_2$ )

$$= \left[ \int_{-\frac{\pi}{\Delta}}^{\frac{\pi}{\Delta}} M(\tau) e^{-i\ell\Delta\tau} d\tau \right] \int_{-\frac{\pi}{\Delta}}^{\frac{\pi}{\Delta}} \int_{-\pi}^{\pi} S(y_1) \tilde{J}_{n-1} \left( \begin{matrix} y_1 - x_2 \Delta \\ x_2 \end{matrix} \right) e^{-imy_1} e^{-i\ell\Delta x_2} dy_1 dx_2$$

(identify  $m_\ell$  from (D-7) )

$$= m_\ell \int_{-\frac{\pi}{\Delta}}^{\frac{\pi}{\Delta}} \int_{-\pi}^{\pi} S(y_1) \tilde{J}_{n-1} \left( \begin{matrix} y_1 - x_2 \Delta \\ x_2 \end{matrix} \right) e^{-imy_1} e^{-i\ell\Delta x_2} dy_1 dx_2$$

(substitute (D-1) for  $\tilde{J}_{n-1}$  and (D-9) for  $S(y_1)$  )

$$= m_\ell \int_{-\frac{\pi}{\Delta}}^{\frac{\pi}{\Delta}} \int_{-\pi}^{\pi} C_0 \sum_{\lambda} S_{\lambda} e^{i\lambda y_1} \sum_j \sum_k \left[ a_{jk}^{n-1} e^{ij(y_1 - x_2 \Delta)} e^{ik\Delta x_2} \right] e^{-imy_1} e^{-i\ell\Delta x_2} dy_1 dx_2$$

(rearrange using Fubini theorem)

$$= C_0 m_\ell \sum_{\lambda} \sum_j \sum_k S_{\lambda} a_{jk}^{n-1} \int_{-\pi}^{\pi} e^{iy_1(j+\lambda-m)} dy_1 \int_{-\frac{\pi}{\Delta}}^{\frac{\pi}{\Delta}} e^{i\Delta x_2(k-\ell-j)} dx_2$$

(observe =  $\nu$  unless  $j = m-\lambda$  and  $k = j+\ell$ )

$$= \frac{4\pi^2}{\Delta} C_0 m_\ell \sum_{\alpha} S_{m-\alpha} a_{\alpha, \ell+\alpha}^{n-1} . \quad (D-10)$$

From the definition (D-2) we observe that  $a_{00}^n = 1$ , since  $\tilde{J}_n$  must have unit total integral. But

$$a_{00}^n = \frac{4\pi^2}{\Delta} C_0 m_0 \sum_{\alpha} S_{-\alpha} a_{\alpha, \alpha}^{n-1} = 1 . \quad (D-11)$$

Accordingly, we have

$$\frac{4\pi^2}{\Delta} C_0 = \frac{1}{m_0 \sum_{\alpha} S_{-\alpha} a_{\alpha, \alpha}^{n-1}} ,$$

So, if we define  $\tilde{m}_\nu = m_\nu/m_0$ , we have finally that

$$a_{m\ell}^n = \frac{\tilde{m}_\ell \sum_{\alpha} S_{m-\alpha} a_{\alpha, \ell+\alpha}^{n-1}}{\sum_{\alpha} S_{-\alpha} a_{\alpha, \alpha}^{n-1}} , \quad (D-12)$$

where

$$\tilde{m}_\ell = \exp \left[ -\frac{q\ell^2\Delta^3}{2} \right] , \quad (D-13)$$

and

$$S_{\alpha} = I_{\alpha} \left( \frac{|z|}{\Delta/r} \right) \exp \{ -1, \emptyset \} . \quad (D-14)$$

Next, we observe that (from (D-2) )

$$a_{-1,0}^n = \int_{-\frac{\pi}{\Delta}}^{\frac{\pi}{\Delta}} \int_{-\pi}^{\pi} e^{-iy_1} \tilde{J}_n \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} dy_1 dy_2 = E[\cos x_1 | z_n] + iE[\sin x_1 | z_n] . \quad (D-15)$$

Therefore, the cyclic estimate (which minimizes  $E[2(1-\cos e)]$  ) is given by

$$\begin{aligned}\hat{x}_1(n|n) &= \tan^{-1} \left\{ E[\sin x_1 | z_n] / E[\cos x_1 | z_n] \right\} \\ &= \tan^{-1} \left\{ \text{Im}(a_{-1,0}^n) / \text{Re}(a_{-1,0}^n) \right\}\end{aligned}\quad (\text{D-16})$$

Using the equations (D-12) - (D-14) and (D-16) we have implemented an example of the Fourier Series filter with  $-5 \leq m \leq 5$  and  $-5 \leq \ell \leq 5$  for a total of  $11 \times 11 = 121$  coefficients. Due to the occurrences of negative mass we discovered that the Fourier Series is not suitable for low-noise situations. On the other hand, for  $N/S = 1$  dB and  $q = 0.1$  we managed to get quite good results. This may be seen in Fig. C-1, where the result of the mean-modulo- $2\pi$ -squared error of the Fourier Series is shown in conjunction with the experimental results for the phase-locked loop (Appendix A), the Hermite Expansion (Appendix B), and the point-mass representation (Appendix C). The nominal Monte Carlo result for the Fourier Series at  $N/S = 1$  dB was 2.65 dB with a 200 Monte Carlo  $3\sigma$  confidence from 2.26 dB to 3.08 dB. This result is equivalent to a nominal improvement over the phase-locked loop of 0.84 dB with  $3\sigma$  confidence from 0.41 dB to 1.23 dB. Also, the difference between Fourier Series performance and the ideal linear was nominally 2.12 dB with confidence interval from 1.73 dB to 2.55 dB.

Although the above Monte Carlo result was consistent with the previous experiments it represents a preliminary result, since we still have not completely isolated a solution to the negative mass dilemma. More results will follow in a later paper.

### Appendix E. A Movie of Conditional Densities

Just as we have demonstrated the value of visual inspection of conditional densities for the tracking problem in the past (Chapter VI), we now illustrate the wealth of information contained in the conditional densities for the phase demodulator. In this appendix, we describe a movie made of cyclic phase densities using the point-mass method described in Appendix C.

The parameters used for the sequence in the movie were as follows:

$$N/S = 0 \text{ dB} , \quad q = 0.1$$

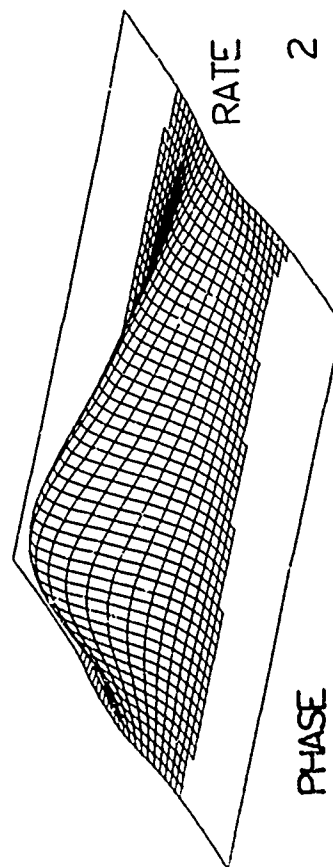
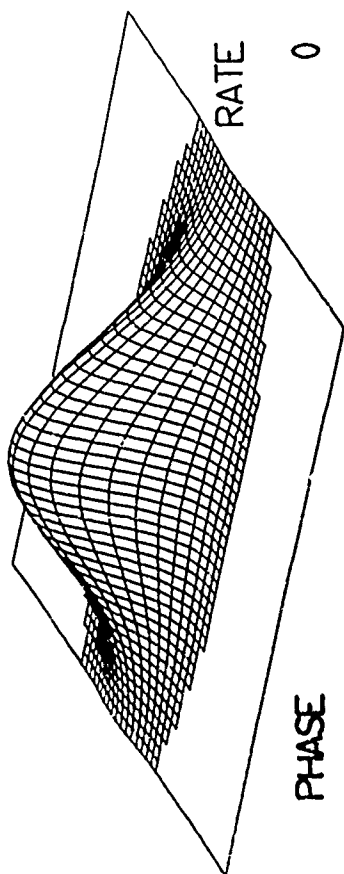
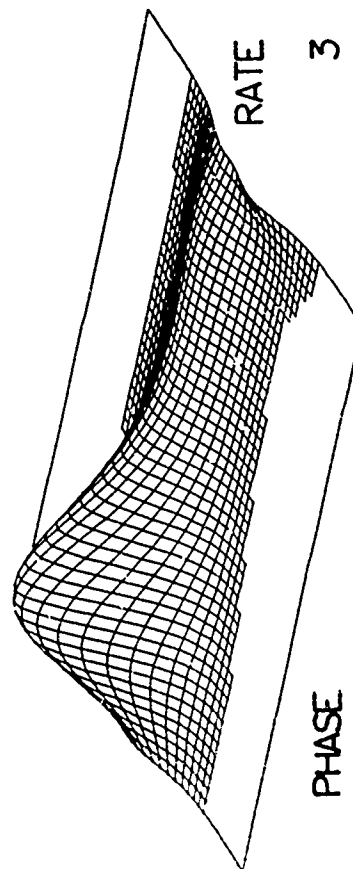
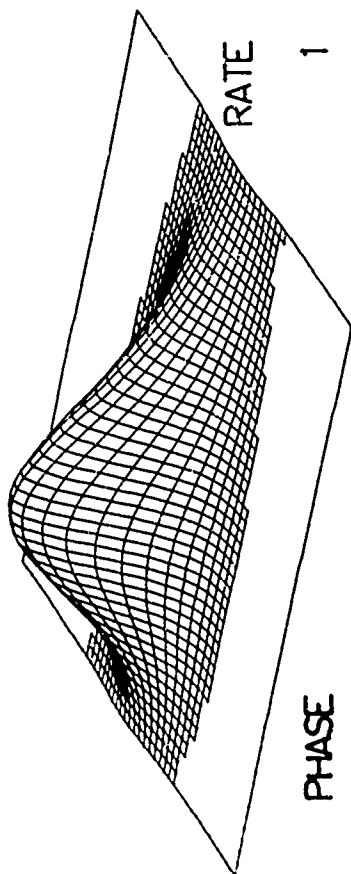
$$\Delta = 0.24$$

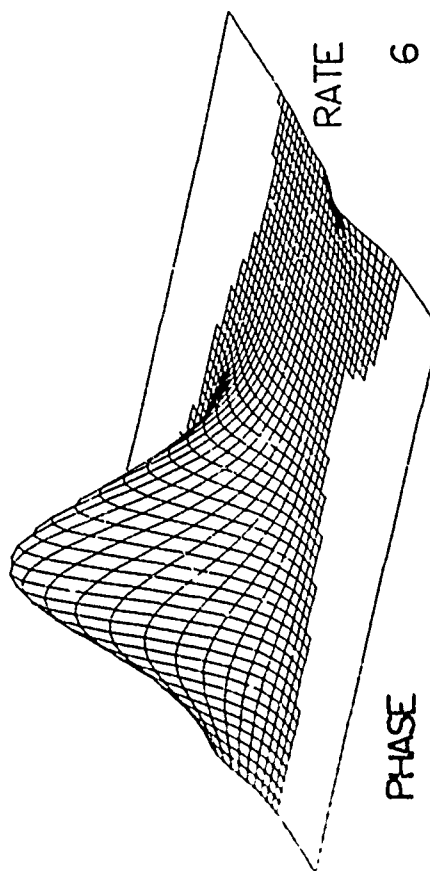
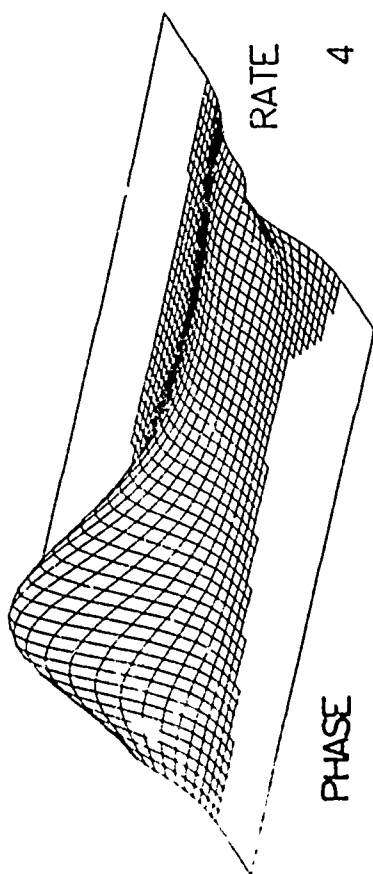
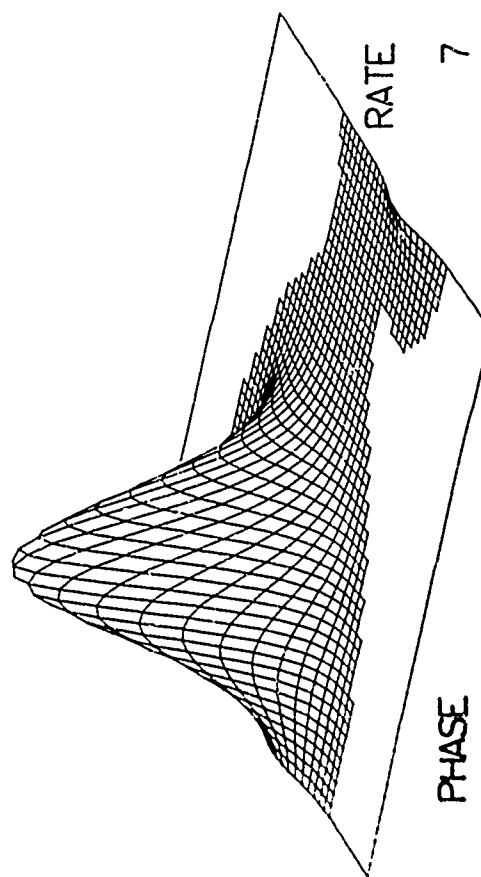
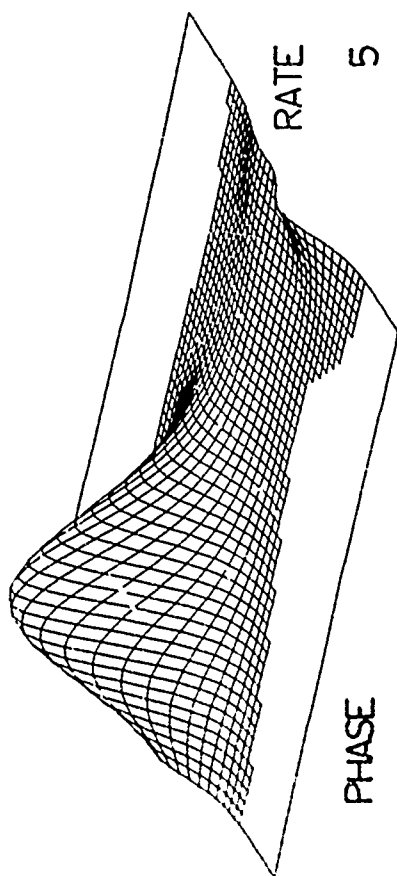
The initial density was chosen with the nominal steady-state value  $P_{11} = 1.85 \text{ dB}$  (1.53), and the grid size was set at 31 points in phase by 155 points in phase-rate. The isometric views of the densities, seen in Fig. E-1, are shown for phase over the entire interval  $[-\pi, \pi)$ , but phase rate is shown over only one third of the interval  $\left[-\frac{\pi}{\Delta}, \frac{\pi}{\Delta}\right)$ . Thus spillover in the phase rate direction is not lost, but merely not shown. In the initial sequence from the movie [3], shown in Fig. E-1, many features may be observed. Cycle slips in phase are accompanied by general turbulence of the density, the appearance of multiple modes and other anomalies. One explanation for the appearance of multiple modes and thus the cycle slips is the occasional major disagreement between the in-line and quadrature measurement components  $z_1$  and  $z_2$ , as a result of the independent noises  $v_1$  and  $v_2$ . The sequence shown in the figure illustrates, though, how recovery is gradually reaccomplished when the measurements agree.

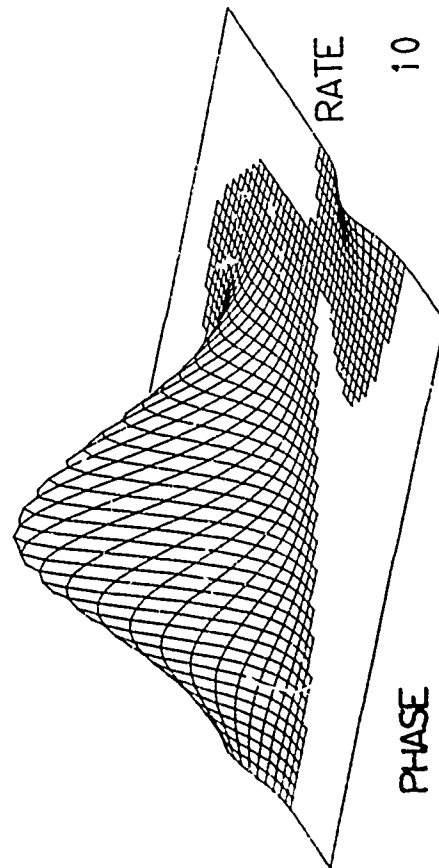
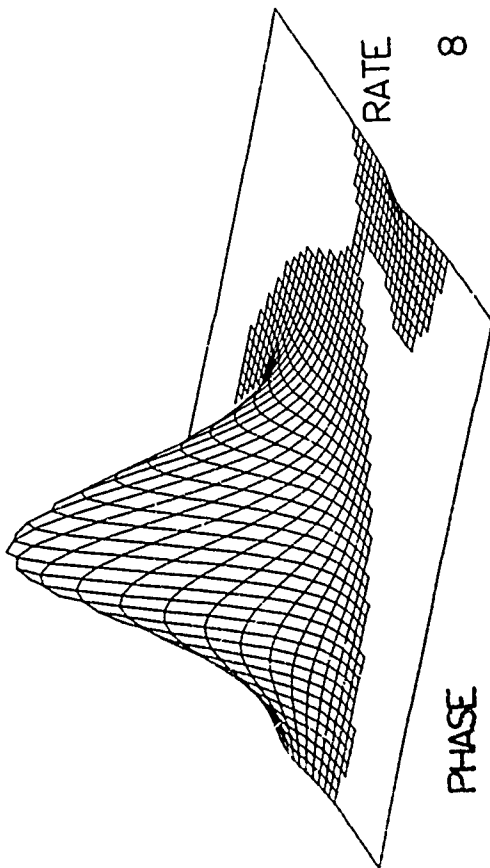
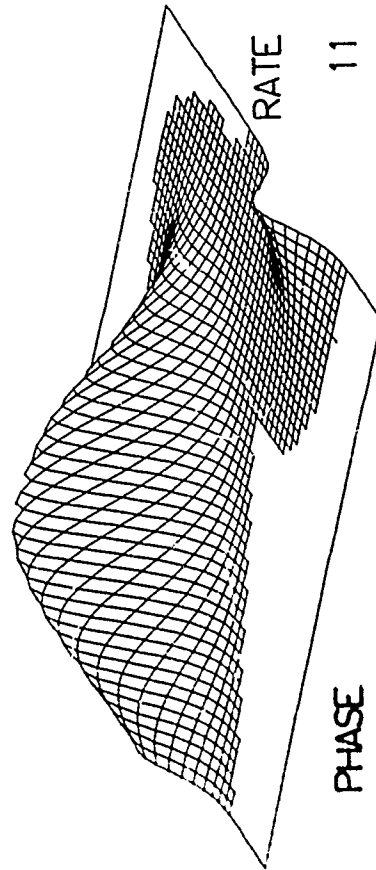
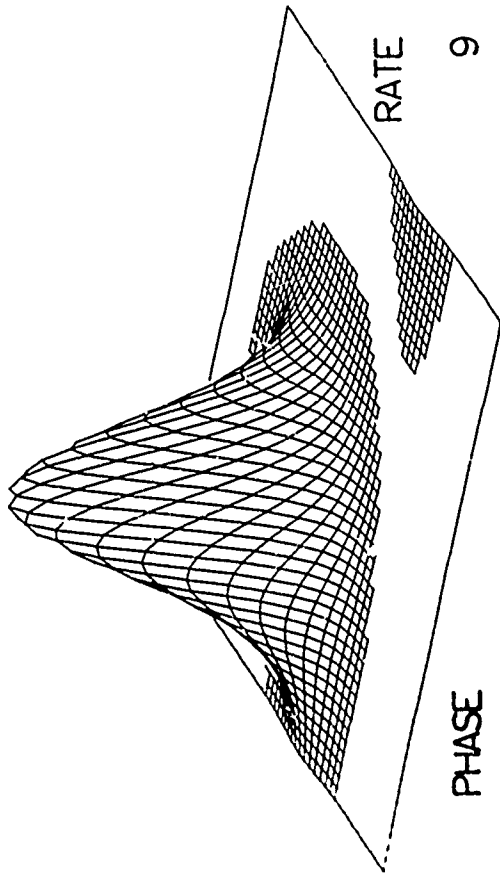
**Preceding page blank**

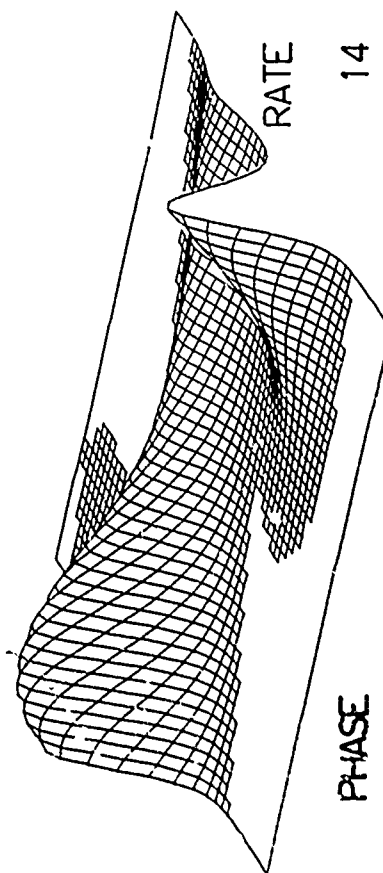
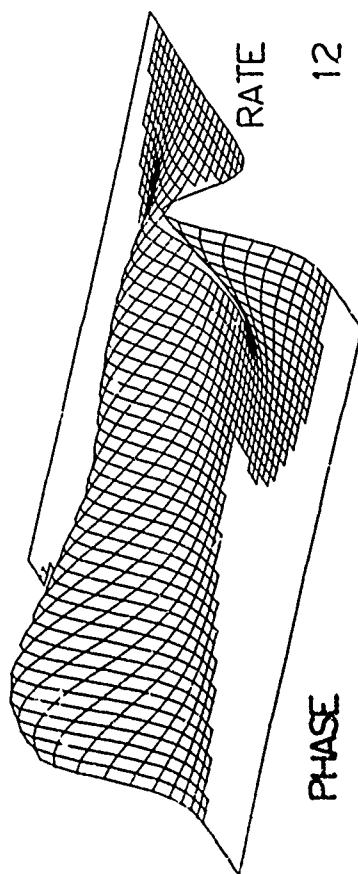
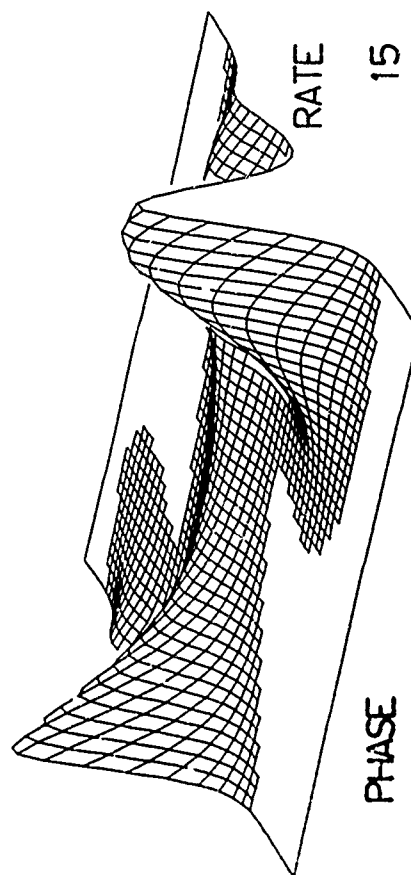
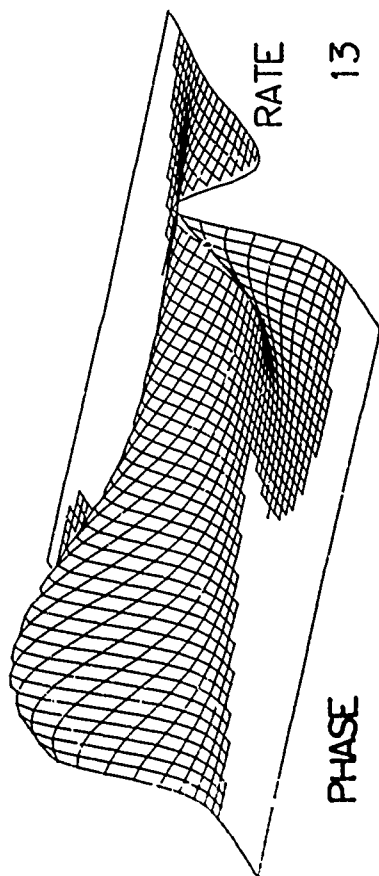
Fig. E-1. A Typical Sample Path of  
Densities Evolving in Time

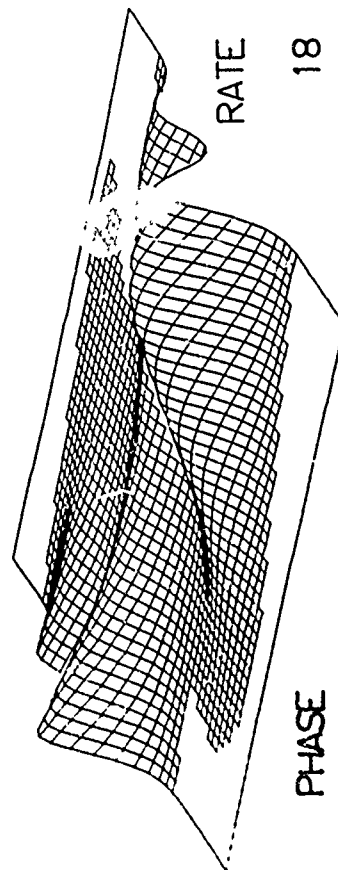
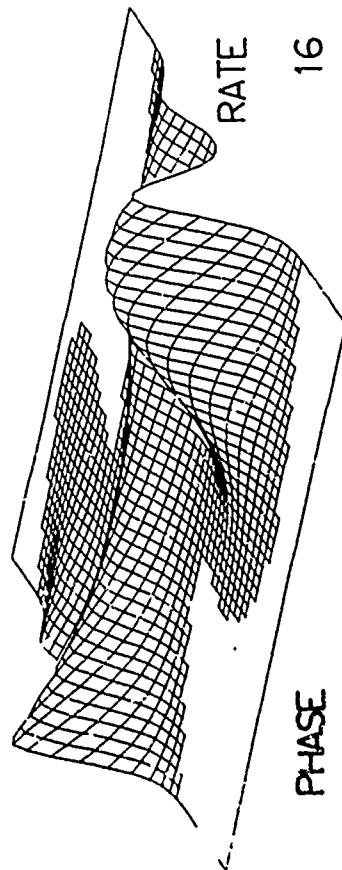
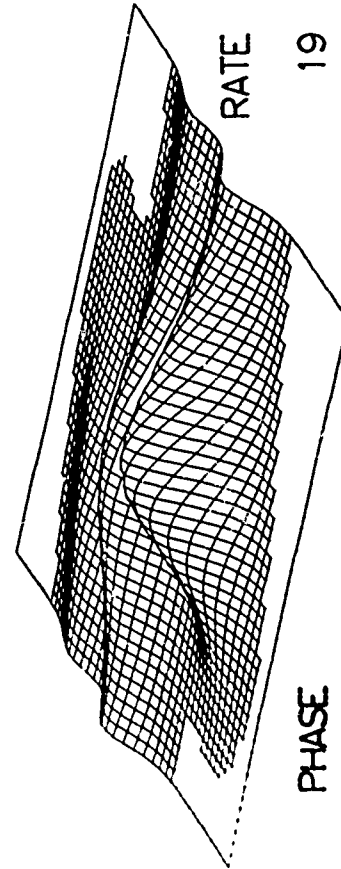
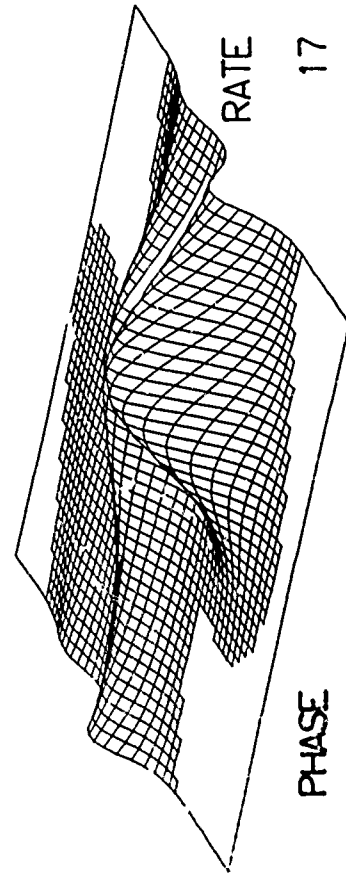
The following 15 pages constitute Fig. E-1. The densities are taken from the initial condition and 59 conditional a posteriori densities.

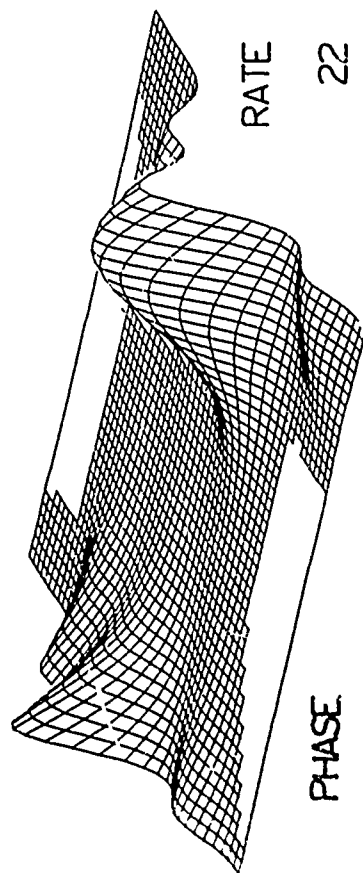
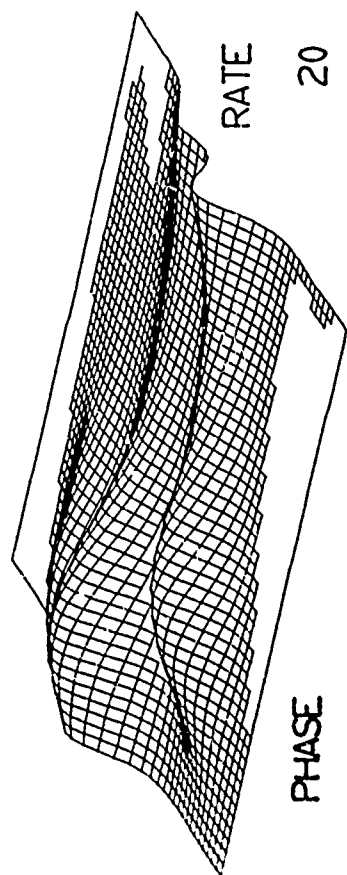
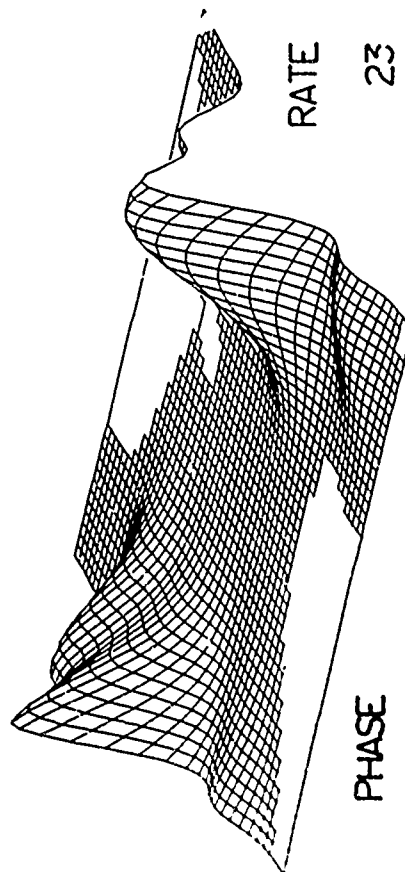
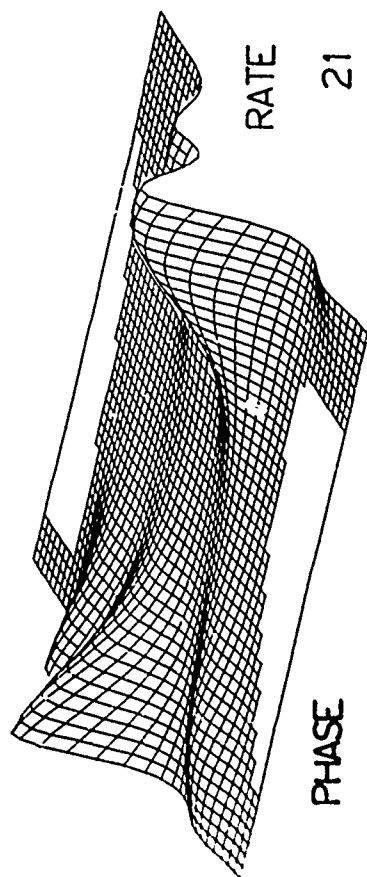


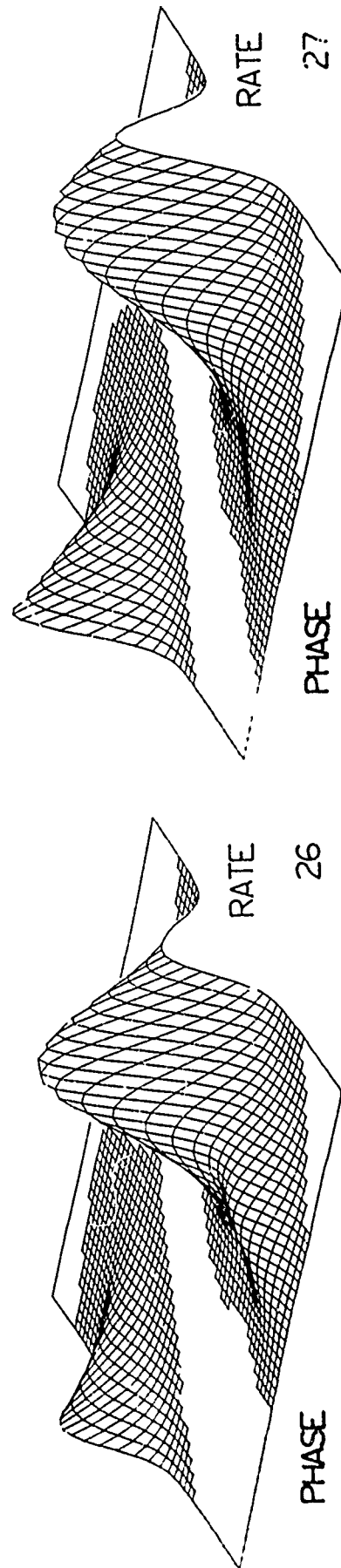
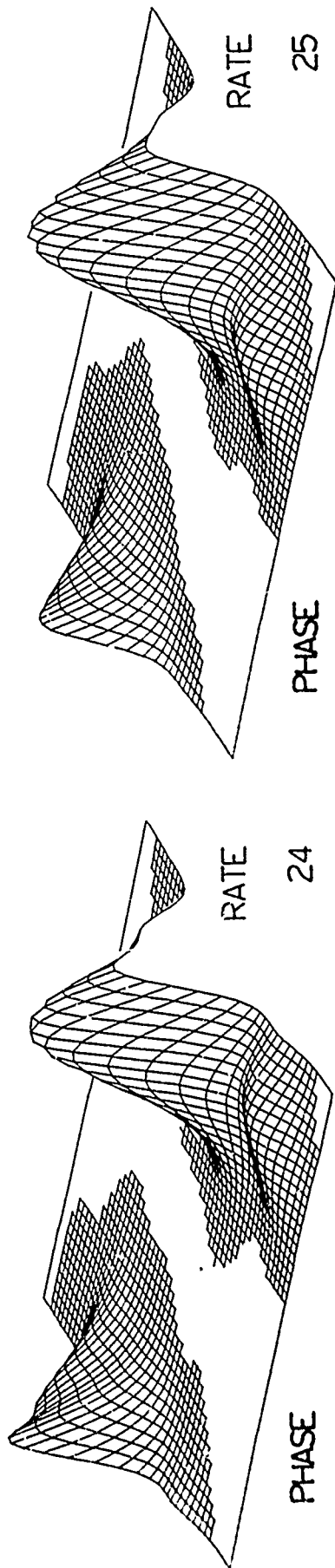


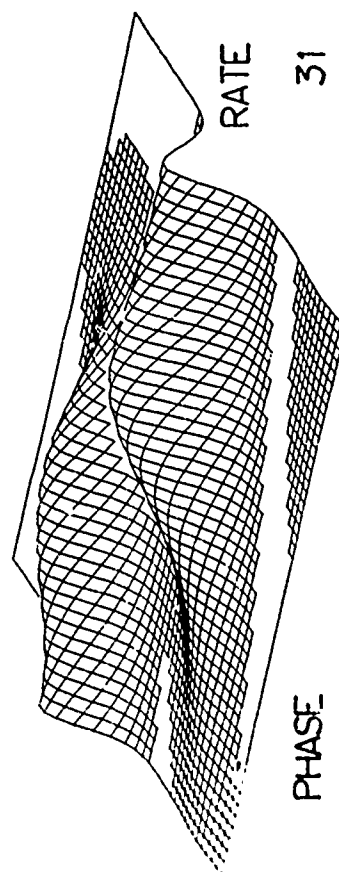
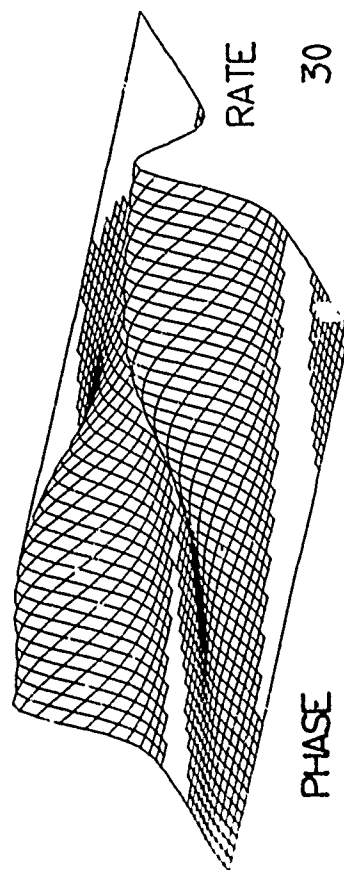
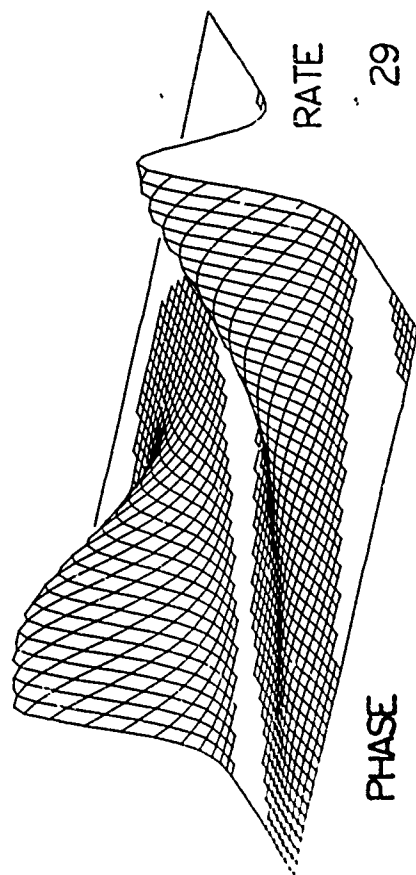
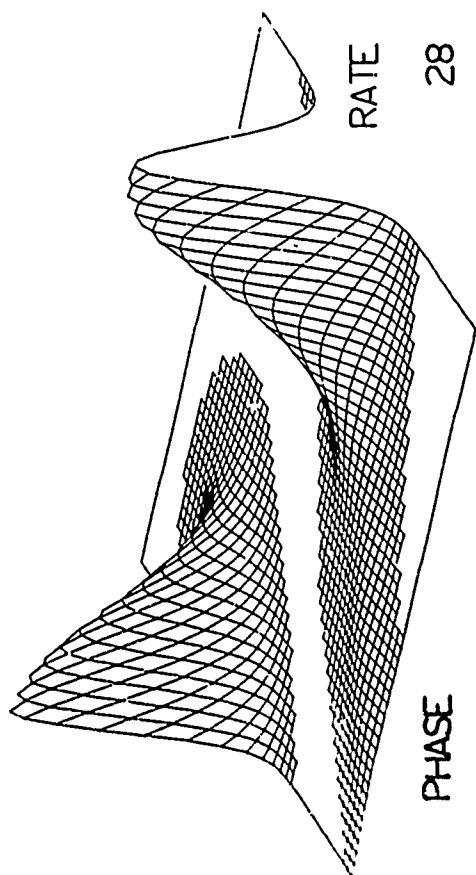


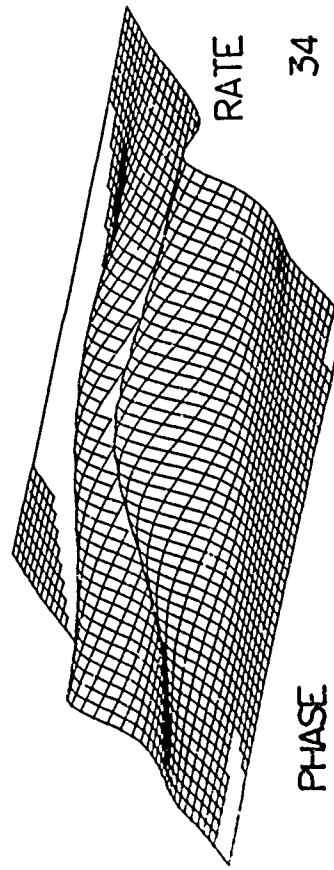
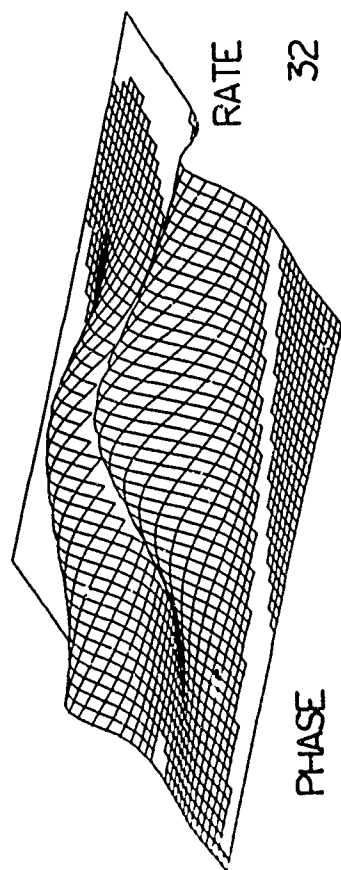
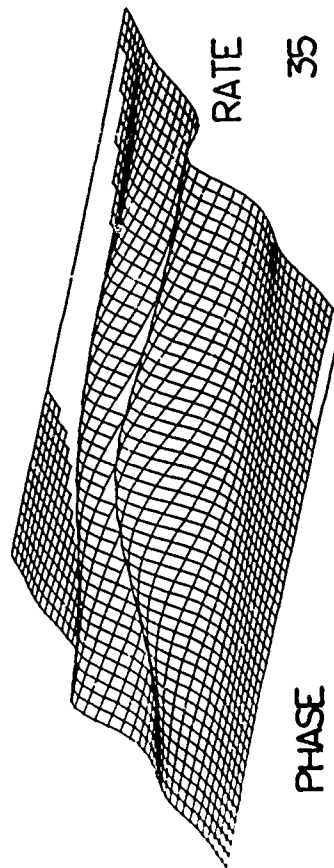
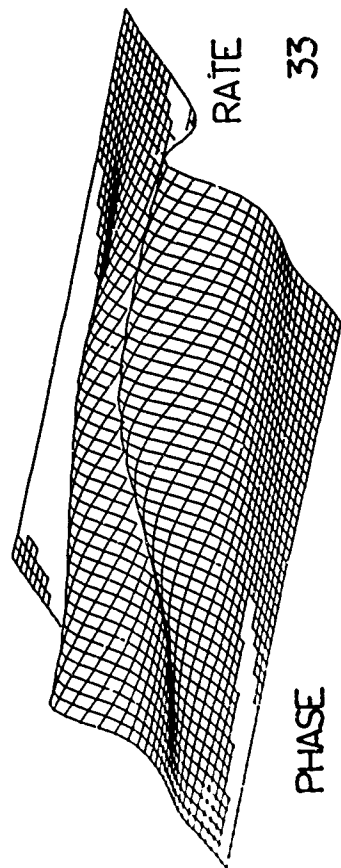


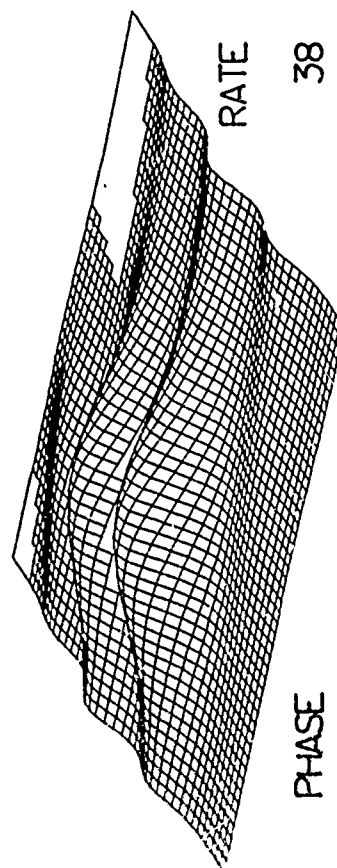
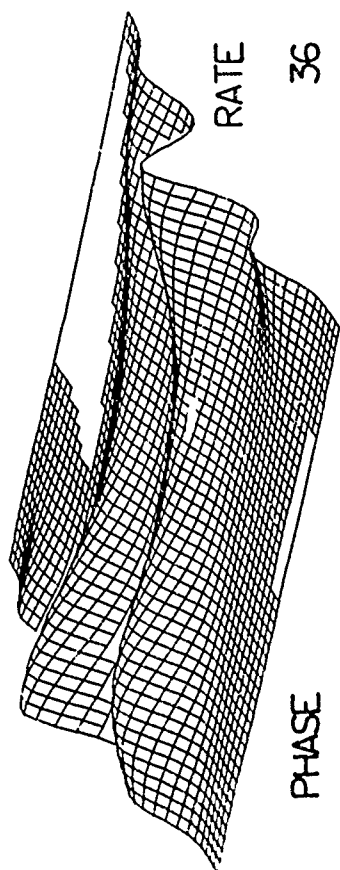
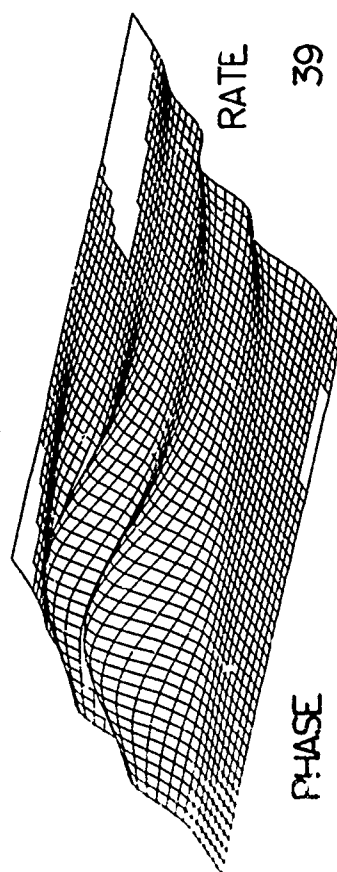
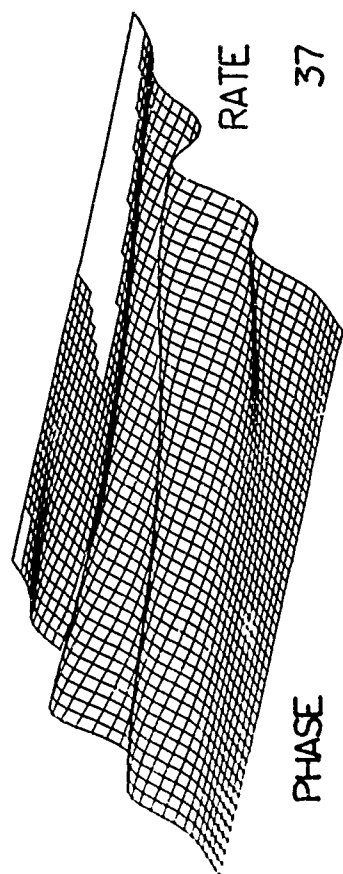


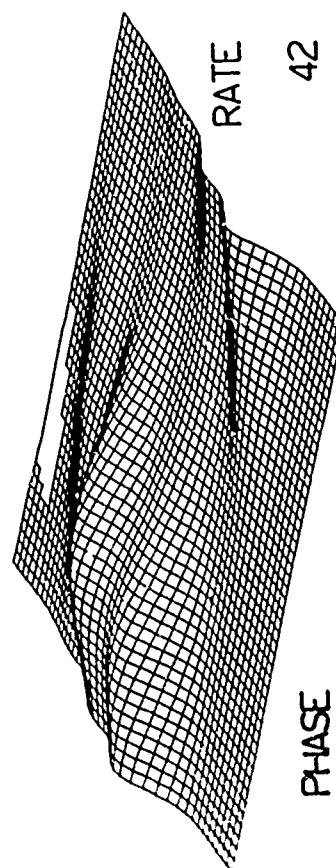
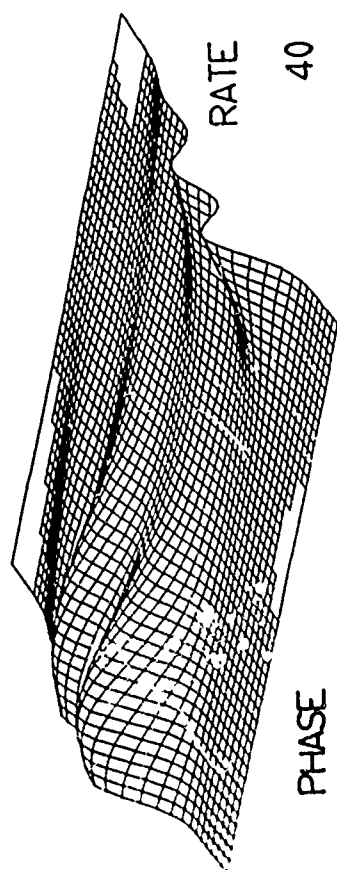
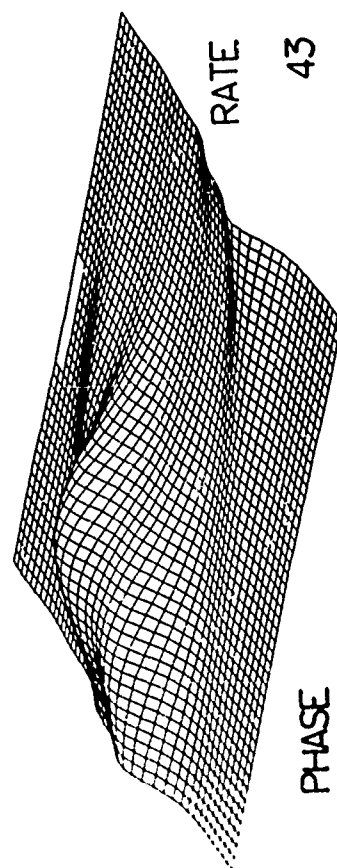
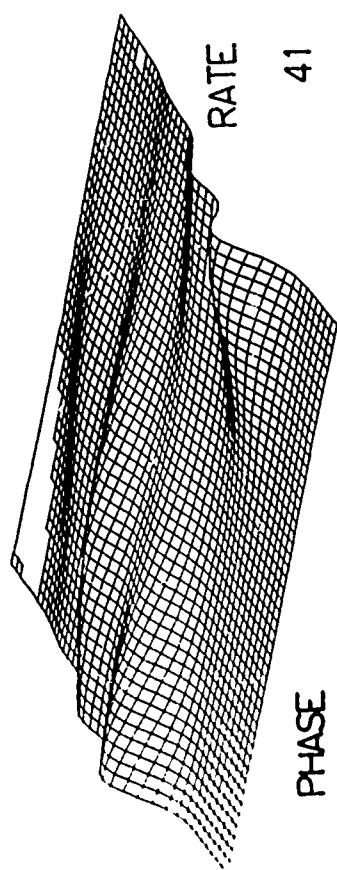


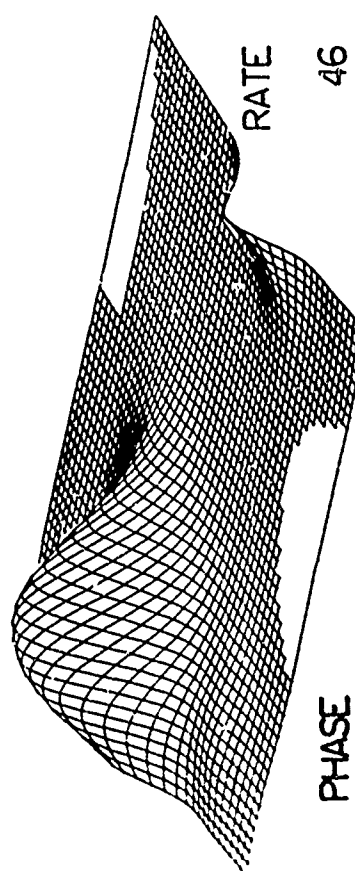
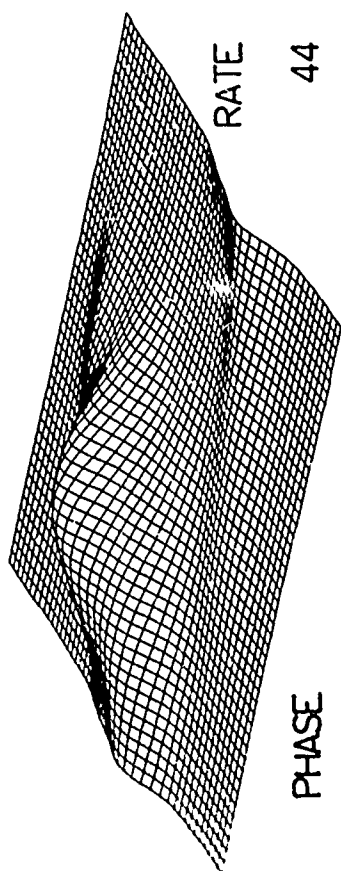
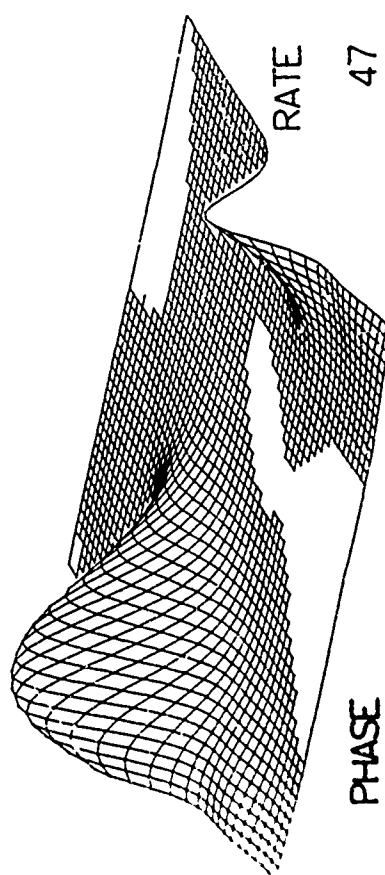
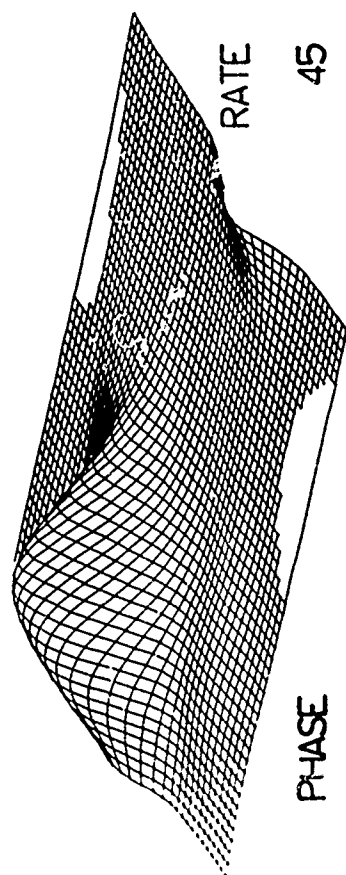


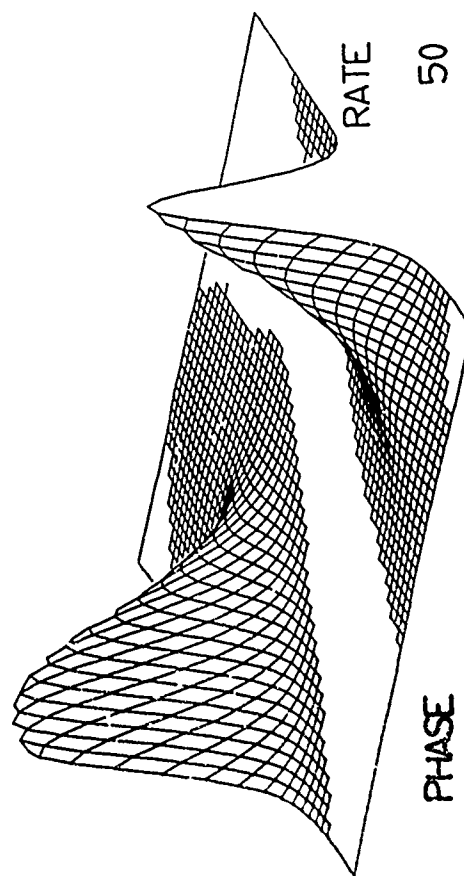
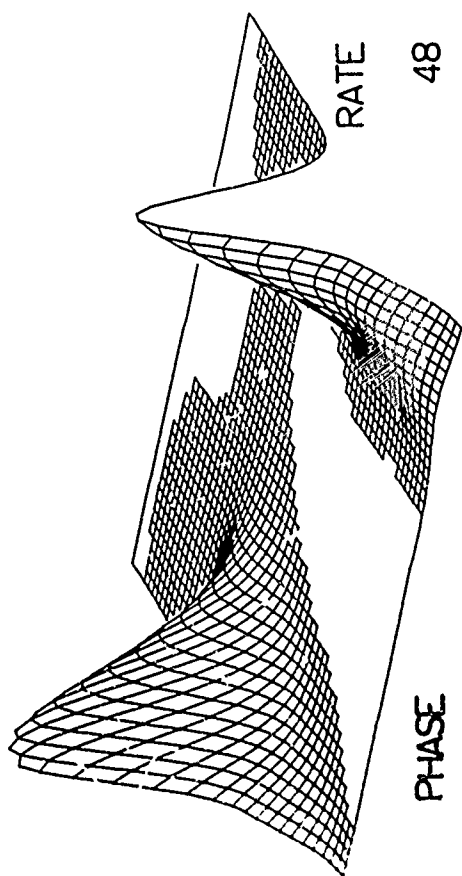
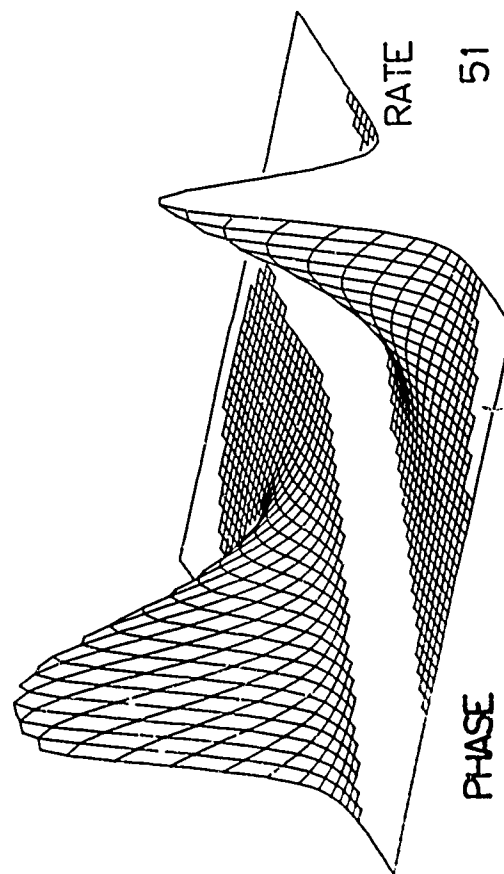
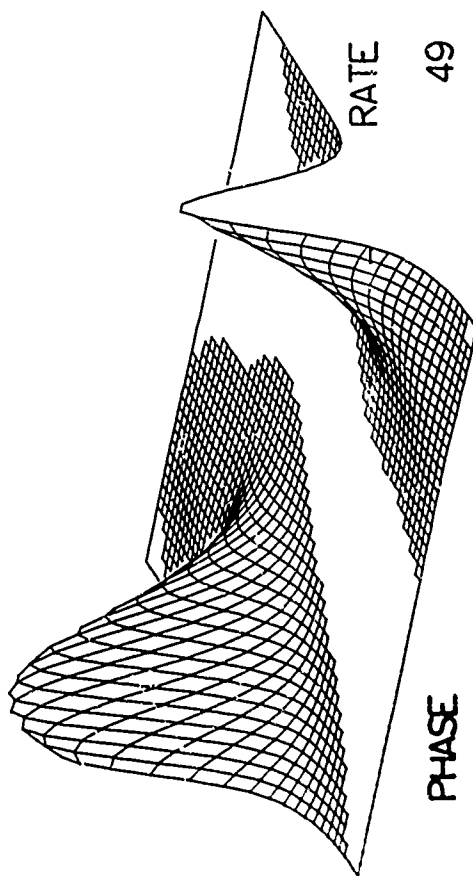


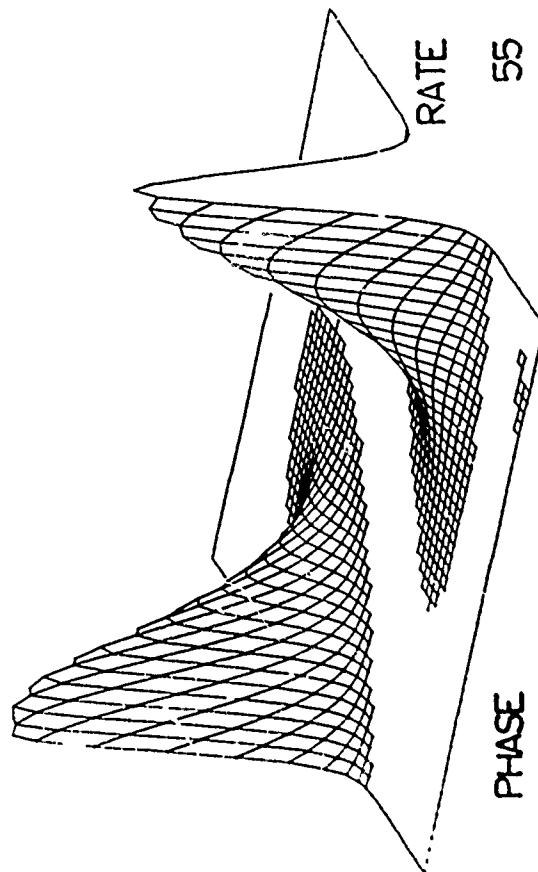
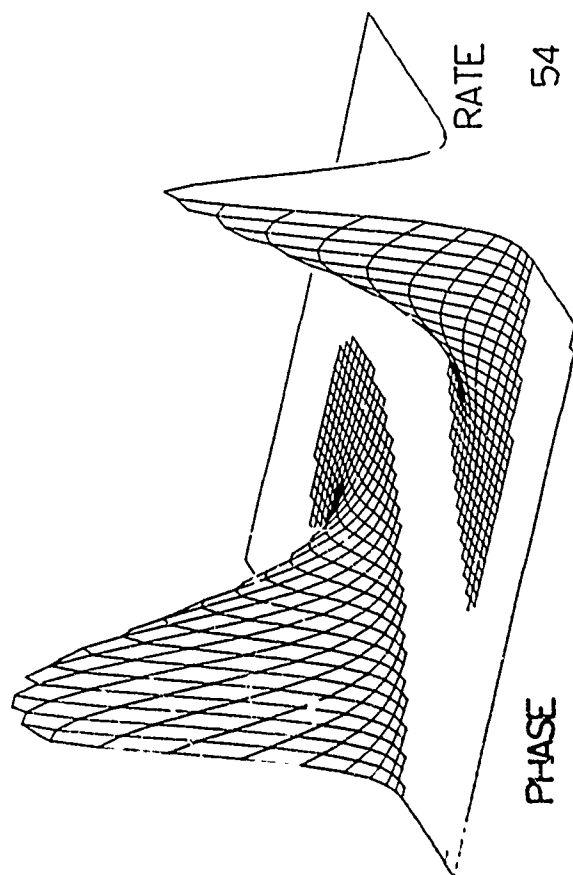
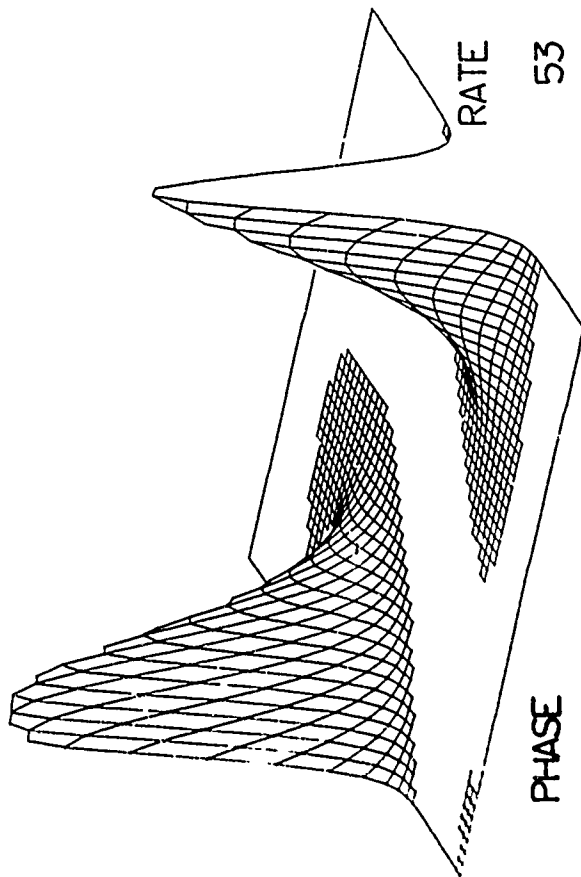
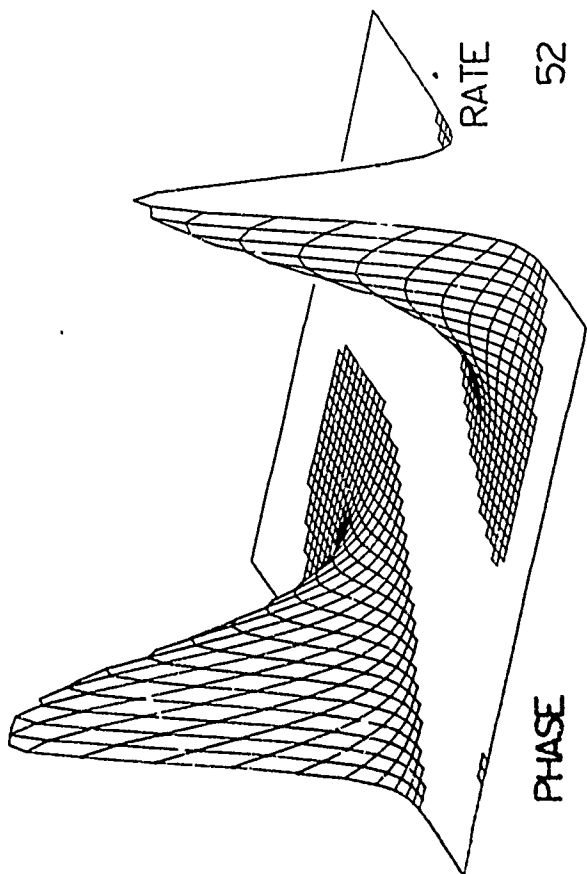


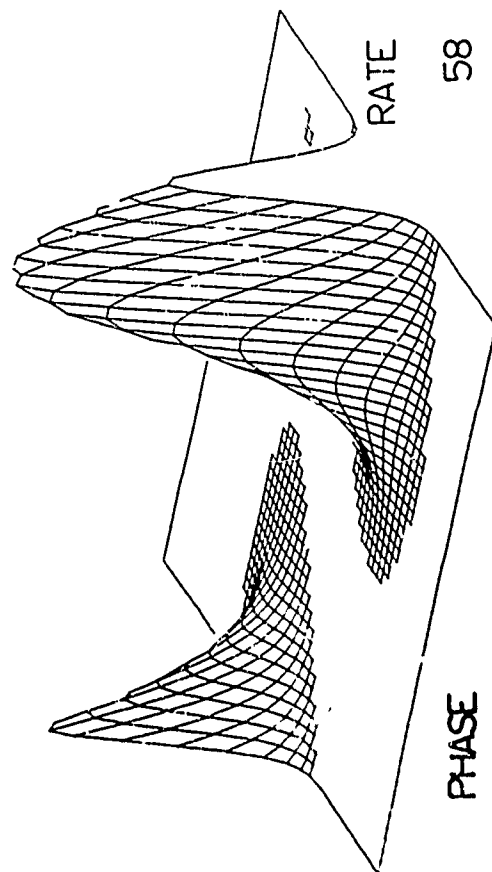
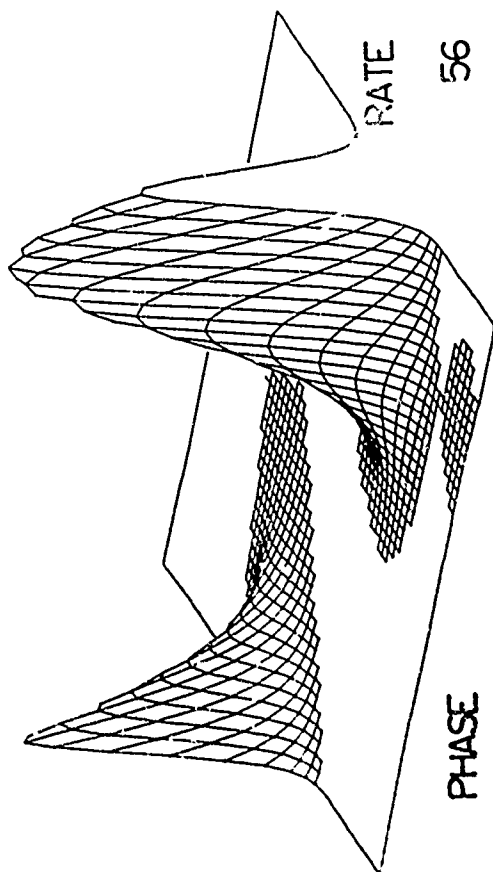
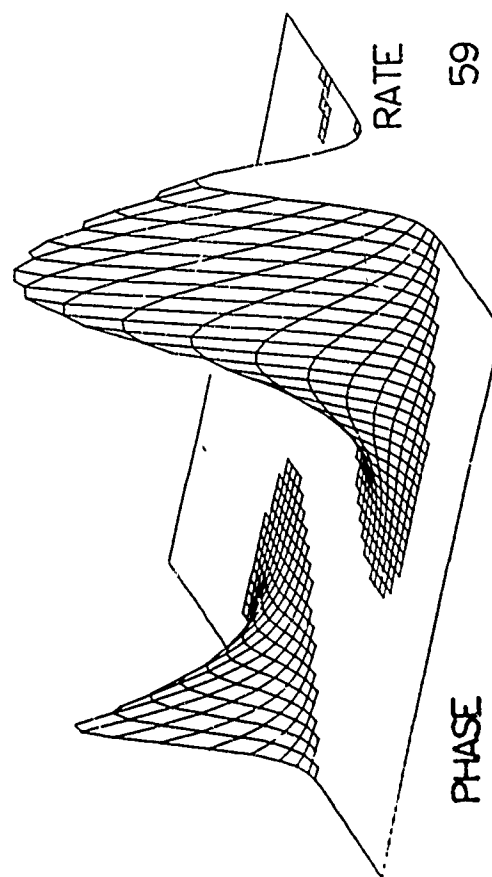
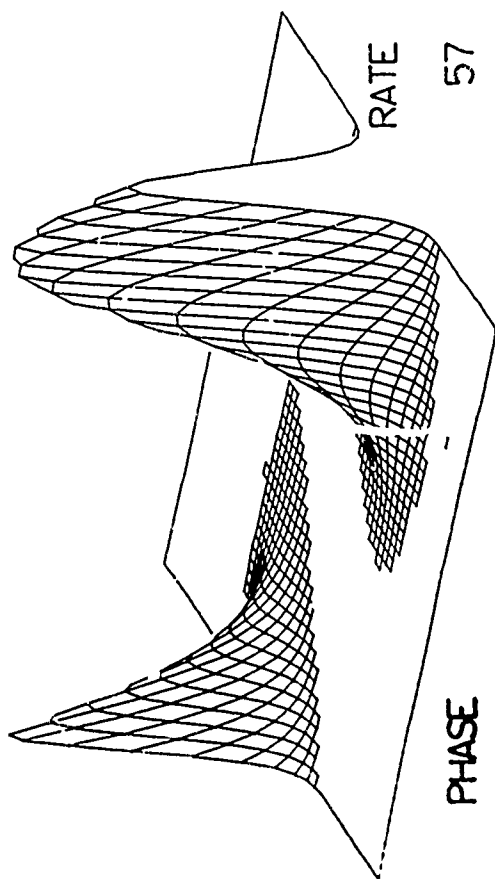












### VIII. Conclusions

The purpose of this report has been to review the extensive research concerning numerical methods of implementing optimal Bayes estimators during the past two to three years. It is anticipated that the contents of this report will represent only a small fraction of the important work on the problem in the next few years. This prediction is based in part on the increasing number of researchers who have turned to the subject during the last two years.

We interpret the effort described herein as a pioneering venture. We have not been able to answer or, in fact, ask every possible question during this research. We have demonstrated the present day feasibility of implementing excellent numerical approximations to nonlinear filters. We have further discussed how this effort can be extended - both in the realm of machine design and in the area of practical engineering application.

Almost any example involving only a few state dimensions would be appropriate for this investigation. We require a small state vector since the evaluation of nonlinear estimators requires costly Monte Carlo simulations - a topic which we covered in great detail in Chapter IV. Another advantage of a small state vector (less than or equal to two dimensions) is the accessibility of the conditional probability densities to visual inspection via computer graphics. We have demonstrated through the use of two movies the important advantage of visual inspection of the conditional densities evolving in time. This of course emphasizes the point that the nonlinear estimate is only a

**Preceding page blank**

byproduct of this research - the principal concern is the conditional density, which we have available for inspection at every sample time.

It would be incorrect to surmise, however, that we do not respect the value of moment series expansions, as exemplified by the extended Kalman-Bucy filter, for example. Rather, we accept past successes at face value - as a testimony to engineering perseverance and diligence and we intend to supplement these efforts with some sound, systematic solutions to the general nonlinear estimation problem which, hopefully, will provide increased understanding of the characteristics of the optimal estimator in a variety of applications.

At the same time, however, it would be pessimistic to consider our methods beyond practicality, or non "real-time". We have noted frequently in the present report that only development time lies between the present ideas and their real-time realizations in digital machines. The fundamental reason for our certainty for this conjecture is the existence of essentially unlimited parallelism in the Bayes-Law computations. We contend furthermore, that pilot experiments on existing parallel machines, yet to be performed, will prove our conjecture that the parallelism in these algorithms may be exploited indefinitely, that perhaps Bayes estimation is truly the first non-academic application of parallel machine architecture.

Of course one important question which we have raised concerns the identification of suitable example problems which illustrate the unique characteristics of nonlinear estimators. It is obvious that our efforts would be partially unnecessary if it were clear at the outset what examples were appropriate. We have presented two important applications for the theory, picked by and large at random, problems which represent

some of the most significant challenges to nonlinear filtering. The first problem concerns passive tracking, a subtle application of filtering to a system involving a singular nonlinearity and all of the potential numerical instabilities that result. The second application involves communication theory in the sense that the limitation on current application of demodulators revolves around the reliability of the phase-locked loop. We have shown that threshold extensions of several dB are possible by routine application of optimal nonlinear estimators. In addition we have demonstrated the advantages of engineering the error loss function and of examining the conditional expected loss for each sample sequence, resulting in considerable performance feedback during the operation of the filter.

We expect, in summary, that our results described in this report will be seminal to an expanding development of experiments and research efforts in the field of synthesizing and evaluating nonlinear estimators. We anticipate, further, that respect for and understanding of numerical approximation methods will continue to develop in both among applied engineers and the mathematicians concerned with nonlinear estimation. It is in this spirit, then, that we offer the report as a manual - an engineer's guide to building nonlinear estimators.

## Bibliography\*

- [ 1] M. Abramowitz and I.A. Stegun, Handbook of Mathematical Functions, Dover, New York, 1965. (IV, VII)
- [ 2] D.L. Alspach, "A Bayesian Approximation Technique for Estimation and Control of Time Discrete Stochastic Systems," Ph.D. Dissertation, University of California, San Diego, 1970. (I, III, V, VI)
- [ 3] D.L. Alspach and H.W. Sorenson, "Approximation of Density Functions by a Sum of Gaussians for Nonlinear Bayesian Estimation," Proc. Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1970, 19-31. (I, VI)
- [ 4] J.L. Baer and D.P. Bovet, "Compilation of Arithmetic Expressions for Parallel Computations," Information Processing 1968, North-Holland Pub. Co., Amsterdam, 1969, 340-346. (V)
- [ 5] R.W. Bass, V.D. Norum, and L. Schwartz, "Optimal Multichannel Non-Linear Filtering," J. Math. Anal. Appl. 16 (1966), 152-164. (I)
- [ 6] W.J. Bouknight, et al. "The Illiac-IV Systems," Proc. IEEE, 60 (1972), 369-388. (V)
- [ 7] R.S. Bucy, "Bayes Theorem and Digital Realizations for Non-Linear Filters," J. Astro. Sci. 17 (1969), 80-94. (I, II, III)
- [ 8] R.S. Bucy, "Realization of Non-Linear Filters," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1971, 51-58. (I)\*\*
- [ 9] R.S. Bucy, "Building and Evaluating Non-Linear Filters," To appear, Proc. Symp. on Appl. Math.; Stochastic Diff. Eqns., Amer. Math. Soc., April 1972. (III, IV, VII)
- [10] R.S. Bucy, R.A. Geesey, and K.D. Senne, "Passive Receiver Design via Nonlinear Filtering Theory," Proc. Third Hawaii International Conf. on System Sciences, Vol I, 1970, 477-480. (I, III, VI)
- [11] R.S. Bucy, C. Hecht, and K.D. Senne, "Optimal Phase Demodulation via Discrete Nonlinear Filtering," Air Force Weapons Laboratory Computer Films No. 72-0401-01, April 1972. (VII)
- [12] R.S. Bucy and P.D. Joseph, Filtering for Stochastic Processes with Applications to Guidance, Wiley Interscience, New York, 1968. (I, II, III, VII)

---

\* Numbers in parentheses after References indicate the Chapters in which the references are cited.

\*\* A copy of this paper appears in Additional Appendix F of this report.

Preceding page blank

- [13] R.S. Bucy, M.J. Merritt, and D.S. Miller, "Hybrid Computer Synthesis of Optimal Discrete Nonlinear Filters," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1971, 59-87. (I, V)\*\*
- [14] R.S. Bucy and K.D. Senne, "Realization of Optimum Discrete-Time Nonlinear Estimators," Proc. Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1970, 6-17. (I)
- [15] R.S. Bucy and K.D. Senne, "Digital Synthesis of Nonlinear Filters," Automatica 7 (1971), 287-298. (I, III, V, VI, VII)
- [16] R.S. Bucy and K.D. Senne, "A Two-Dimensional Passive Ranging Experiment using Optimal Nonlinear Filtering," Air Force Weapons Laboratories Computer Film No. 71-0330-02, March 1971. (VI)
- [17] J.L. Center, "Practical Nonlinear Filtering of Discrete Observations by Generalized Least Squares Approximation of the Conditional Probability Distribution," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1971, 88-99. (I, III, V)
- [18] Control Data 6400/6500/6600 Computer Systems Reference Manual, CDC Publication 60100000, St. Paul, 1968, Chaps 2 and 3. (V)
- [19] H. Cramer, Mathematical Methods of Statistics, Princeton University Press, Princeton, 1951, pp 416-451. (IV)
- [20] W.F. Denham and S. Pines, "Sequential Estimation when Measurement Function Nonlinearity is Comparable to Measurement Error," AIAA J. 4 (1966) 1071-1076. (VI)
- [21] J.L. Doob, "Heuristic Approach to the Kolmogorov-Smirnov Theorems," Annals of Mathematical Statistics, 20 (1949), 393-403. (IV)
- [22] R.J.P. deFigueiredo and Y.G. Jan, "Spline Filters," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1971, 88-99. (I, III, VI)
- [23] C. Hecht, "Digital Realization of Non-Linear Filters," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1971, 152-158. (I, V)\*\*
- [24] C. Hecht, "Synthesis and Realization of Nonlinear Filters," Ph.D. Dissertation, University of Southern California, 1972. (I, III, VII)
- [25] F.B. Hildebrand, Introduction to Numerical Analysis, McGraw Hill, New York, 1956. (III)
- [26] Y.G. Jan, Ph.D. Dissertation, Rice University, 1971. (I, III)
- [27] A.H. Jazwinski, Stochastic Processes and Filtering Theory, Academic Press, New York, 1970. (I)

- [28] R.M. Keller, "On Maximally Parallel Schemata," IEEE Conf. Rec. 11th Annual Symp. on Switching Theory and Automata Theory, 1970, 32-50 (V)
- [29] R.E. Larson and E. Tse, "Modal Estimation and Parallel Computers," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, 1971, 188-197. (V)
- [30] J.T. Lo, "Finite Dimensional Sensor Orbits and Nonlinear Filtering," Ph.D. Dissertation, University of Southern California, 1969. (III)
- [31] A.J. Mallinckrodt, R.S. Bucy, and S.Y. Cheng, "Final Project Report for a Design Study for an Optimal Non-Linear Receiver/Demodulator," NASA Contract NAS5-10789, Goddard Space Flight Center, Maryland, 1970. (I, VII)
- [32] F.J. Massey, Jr., "A Note on the Estimation of a Distribution Function by Confidence Limits," Annals of Mathematical Statistics, 21 (1950), 116-119. (IV)
- [33] W.C. Meilander, "The Associative Processor in Aircraft Collision Prediction," NAECON Proc., 1968. (V)
- [34] D.S. Miller, "Hybrid Synthesis of Optimal Discrete Nonlinear Filters," Ph.D. Dissertation, University of Southern California, 1971. (I, V)
- [35] K.D. Senne, "Bayes Law Implementation: Optimal Discrete-Time Phase Estimation," Proc. SWIEEECO Conf., Dallas, April 1972. (I)
- [36] K. D. Senne, "Computer Experiments with Nonlinear Filters," Proc. Second Symp. on Nonlinear Estimation and Its Applications, San Diego, 1971, 314-324 (Misprinted - see Additional Appendices of this report for corrected version). (VI)\*\*
- [37] K.D. Senne and R.S. Bucy, "Digital Realization of Optimal Discrete-Time Nonlinear Estimators," Proc. Fourth Annual Princeton Conf. on System Sciences, Princeton, March 1970, 280-284. (I)
- [38] H.W. Sorenson and D.L. Alspach, "Recursive Bayesian Estimation using Gaussian Sums," Automatica, 7 (1971), 465-479. (I)
- [39] H.W. Sorenson and A.R. Stubberud, "Non-Linear Filtering by Approximation of the A Posteriori Density," International J. Control, 8 (1968), 33-51. (I)
- [40] K. Srinivasan, "State Estimation by Orthogonal Expansion of Probability Distributions," IEEE Trans. Auto. Control, AC-15 (1970), 3-10. (I)
- [41] E. Tse, "Parallel Computation of the Conditional Mean State Estimate for Nonlinear Systems," Proc. Second Symp. on Nonlinear Estimation Theory and Its Applications, San Diego, Sept. 1971, 385-394. (I, V)

- [42] A.J. Viterbi, Principles of Coherent Communication, McGraw Hill, New York, 1966. (VII)
- [43] H.L. Weinert and T. Kailath, "Recursive Spline Interpolation and Least-Squares Estimation," submitted to Amer. Math. Soc., 1971. (I, III)
- [44] N. Wiener, The Fourier Integral and Certain of Its Applications, Cambridge, Cambridge University Press, 1933 (Also: New York, Dover, 1958).

Resumes of the Authors

RICHARD S. BUCY

B.S. Mathematics, 1957, Massachusetts Institute of Technology

Ph.D. Mathematical Statistics, 1963, University of California,  
Berkeley

1955 : Research U.S. Census Bureau, Statistical  
Standards Division, Suitland, Maryland

1957-60 : Research Mathematician, Johns Hopkins Applied  
Physics Laboratory, Howard County, Maryland

1960-61 : Research Mathematician, Research Institute for  
Advanced Study, A Division of the Martin Company,  
Baltimore, Maryland

1961-63 : Teaching Assistant, University of California,  
Berkeley, California

1963-64 : Research Mathematician, Research Institute for  
Advanced Study, Baltimore, Maryland

1963-64 : Consultant, the RAND Corporation, Santa Monica,  
California

1964-65 : Assistant Professor, Mathematics Department,  
University of Maryland, College Park, Maryland

1964-67 : Consultant, the RAND Corporation, Santa Monica,  
California

1965-66 : Associate Professor, Department of Aerospace  
Engineering Sciences, University of Colorado,  
Boulder, Colorado

1965-66 : Consultant, Martin Marietta Corporation, Denver,  
Colorado

1966-70 : Associate Professor, Departments of Mathematics  
and Aerospace Engineering, University of Southern  
California, Los Angeles, California

1966- : Consultant, TRW Systems, Redondo Beach, California

1967, Summer : Consultant, D.V.L. Institute for Guidance and  
Control, Oberpfaffenhofen, Germany

1968- : Consultant, Electrac Inc.

1969- : Consultant, The Aerospace Corporation, Los Angeles,  
California

- 1970- : Professor, Departments of Mathematics and  
Aerospace Engineering, University of Southern  
California
- 1971- : Editor "Stochastics" Journal of Gordon and Breach

Publications

0. 32 Internal Memoranda, Johns Hopkins Applied Physics Laboratory, Howard County, Maryland.
1. MR-2848<sup>(1)</sup>, "Extreme Positive Definite Functions", (with G. Maltese), Jour. Math. Anal. Appl. 12, 1965, 371-377.
2. MR-5361, "Recurrent Sets", Ann. Math. Stat. 36, 1965, 535-545.
3. MR-8414, "Stability and Positive Supermartingales", Journ. Diff. Equat. 1, 1965, 151-155.
4. MR-5664, "Asymptotic Control Theory", (with R.E. Bellman), J. SIAM Control 2, 1964, 11-18.
5. "An Example of a Transient Set Which has the Property That the Expected Number of Visits is Infinite", The RAND Corporation, RM-3864-PR 1963, 1-4.
6. MR-3349, "A Representation Theorem for Positive Functionals on an Involution Algebra", (with G. Maltese), Math. Annalen, 162, 1966, 364-367.
7. MR-4785, "New Results in Asymptotic Control Theory", J. SIAM Control 4, 1966, 397-402.
8. MR-3076, "New Results in Linear Filtering and Prediction Theory", (with R.E. Kalman), J. Basic Eng., Series D, 1961, 95-108.
9. MR-4940, "Adaptive Finite Time Filtering", (with J.W. Follin), IRE Trans. Auto. Control AC-7, 1962, 10-19.
10. "Nonlinear Filtering", IEEE Trans. Auto. Control, 1965, 198.
11. MR-6479, "Optimal Filtering For Correlated Noise", J. Math. Anal. Appl. 20, 1967, 1-8.
12. "Lectures on Queing Theory", (with R. W. Bass), Department of Aerospace Engineering, University of Colorado, Boulder, Colorado, 1965.
13. "Spinors", Undergraduate Thesis, Mathematics Department, M.I.T., Cambridge, 1957.

(1) Reference to Mathematical Reviews.

14. MR-2846 Filtering for Stochastic Processes with Applications to Guidance, (with P. D. Joseph), Tracts in Pure and Applied Math., Interscience, New York, 1968.
15. "Hunt Revisited", Statistics Department, University of California, Berkeley, 1962.
16. "Discrete Filtering Theory", T. R. W. Internal Report, Redondo Beach, California, 1966.
17. "Comment on 'The Kalman Filter and Nonlinear Estimates of Multivariate Normal Process'", (with R. E. Kalman), IEEE Trans. Auto. Control, 1965, 118.
18. MR-543, "Linear Positive Machines", J. Computer and Systems Sciences 1, 1967, 24-28.
19. "Bayes Theorem and Digital Realizations of Nonlinear Filters", J.A.A.S., 2, 1969, 80-94.
20. "Fundamental Study of Adaptive Control Systems", (with R. E. Kalman and T. Englar), Tech. Report No. ASD-TR-61-27, Vol. 1 and II, 1962 and 1964.
21. MR-5482, "Global Theory of the Riccati Equation", J. Comp. Sys. Sci. 1, 1967, 349-361.
22. MR-493, "Two Point Boundary Value Problems of Linear Hamiltonian Systems", SIAM J. Appl. Math. 15, 1967, 1385-1389.
23. MR-5485, "Canonical Forms for Multivariate Systems", IEEE Trans. Auto. Control AC-13, 1968, 567-569.
24. Article in Contributions to Functional Analysis, (with G. Maltese), Springer-Verlag, Berlin, 1966.
25. MR-3073, "Recent Results in Linear and Nonlinear Filtering", Stochastic Problems in Control, A. S. M. E., New York, 1968.
26. "Linear and Nonlinear Filtering Theory", The RAND Corporation, RM-4735-PR, 1965.
27. MR-2000 "Generalizations of a Theorem of Dolezal", (with L. M. Silverman), J. Computer Sciences, 4, 1971, 334-339.
28. MR-2418, "Linear and Nonlinear Filtering Theory", Proc. Ninth J.A.C.C. Ann Arbor, Michigan, 1968.

RICHARD S. BUCY (Cont'd)

29. "Passive Receiver Design via Nonlinear Filtering Theory", (with R. Geesey and K. Senne), Proc. Third Hawaii Conference on Systems Science, Honolulu, 1970.
30. "Linear and Nonlinear Filtering", Invited Tutorial Review Paper, Proceedings of IEEE, 58, 6, 1970, 854-864.
31. MR-1720 "Correlated Noise Filtering and Invariant Directions for the Riccati Equation", (with D. Rapoport and L. Silverman), IEEE Trans. Auto. Control, 5, 1970, 535-540.
32. "Filtering and It's Applications", Proceedings Third I.F.A.C. Control In Space, Toulouse, France, 1970.
33. MR-2841 "Recent Results in Correlated Noise Filtering", (with D. Rappaport), Proceedings Allerton Conference on Circuit and Systems Theory, 1968.
34. "Singular Problems in Filtering and Control", (with D. Rappaport and L. Silverman), Proceedings Third Hawaii Conference on Systems Sciences, Honolulu, 1970.
35. "Synthesis of Nonlinear Filters", Proceedings 4th Princeton Conference on Information Sciences and Systems, 1970.
36. "Digital Realizations of Optimal Discrete-Time Nonlinear Estimators", (with K. Senne), Proceedings 4th Princeton Conference on Information Sciences and Systems, 1970.
37. "Über die Anzahl der Parameter von Mehrgrossensystem", Regelungstechnik und Process-Datenverarbeitung, 10, 1970, 451-452 (with J. Ackermann).
38. "Canonical Minimal Realizatio. of a Matrix of Impulse Response Sequences" (with J. Ackermann), Information and Control, 19, 1971, 224-231.
39. "A Priori Bounds for the Riccati Equation", to appear Proc. 6th Berkeley Symp. on Mathematical Statistics and Probability, 1970.
40. "Digital Realizations of Non-linear Filters", (with K. Senne), Automatica, 7, 1971, 287-298.
41. "Digital Realizations of Optimal Discrete Time Non-linear Filters", (with K. Senne), Proc. Symposium on Non-linear Filtering, September 1970. Western Periodicals.
42. "A Design Study for an Optimal Non-linear Receiver/Demodulator", Final Report NASA Goddard Space Flight Center, Contract # NAS 5-10789, August 31, 1970 (with A. J. Mallinckrodt and S. Y. Cheng).

RICHARD S. BUCY (Cont'd)

43. "The Riccati Equation and its Bounds" to appear Journal of Computers and System Sciences.
44. "Realization of Non-linear Filters" Proceedings 2nd Sym. on non-linear estimation, La Jolla, California Sept. 13-15, 1971.
45. "Hybrid Computer Synthesis of Optimal Discrete non-linear filters" (with D. S. Miller and M. J. Merritt) Proc. 2nd Sym. on non-linear Estimation, La Jolla, Calif. Sept. 13-15, 1971 to appear Stochastics.
46. A note on "Bounds for the Solutions of the Riccati Equation" I.E.E.E. Automatic Control 17, 1972, 179.
47. "A Negative Definite Equilibrium and its Induced of Global Existence for the Riccati Equation", (With J. Rodriguez-Canabal) to appear, S.I.A.M. Jour. of Math. Analysis, 1972.
48. "An Optimal Phase Demodulator", (with A. J. Mallinckrodt) to appear Stochastics 1972.
49. "Building and Evaluating Non-Linear Filters", to appear Proceeding Symposium on Applied Math, Stochastic Differential Equations, presented April 30, 1972, Am. Math. Soc. Spring Meeting, New York.

CALVIN HECHT

B.S. Mechanical Engineering, 1951, Newark College of Engineering  
 M.S. Engineering, 1955, Massachusetts Institute of Technology  
 Ph.D. Aerospace Engineering, 1972, University of Southern California

1951-54 : Research and Development Engineer, Naval Air  
 1955-56 : Development Center, Johnsville, Pennsylvania  
  
 1956-59 : Flight Test Engineer, General Dynamics Corporation,  
 San Diego, California and Cape Canaveral, Florida  
  
 1959-62 : Engineering Supervisor, North American Rockwell,  
 Autonetics Division, Anaheim, California  
  
 1962-69 : Member of the Technical Staff, North American  
 Rockwell, Space Division, Downey, California  
  
 1970-72 : Research Assistant, Department of Aerospace  
 Engineering, University of Southern California  
  
 1970-71 : Teaching Assistant, Department of Aerospace  
 Engineering, University of Southern California  
  
 1972 : Member of the Technical Staff, TRW Systems,  
 Redondo Beach, California

Publications and Patents

1. "Long Range Vectoring Using the WARRIOR System," M.S. Thesis;  
 May 1955, Massachusetts Institute of Technology, Cambridge,  
 Massachusetts.
2. "The Accuracy Limit of Area Correlation Guidance," (with others).  
 Paper presented at IEEE Aerospace Systems Conference, Seattle,  
 Washington, July 11, 1966.
3. Patent disclosure (PF64M163) for a logic correlator for guidance  
 sensing, September 15, 1964.
4. "Digital Realization of Nonlinear Filters," Proceedings of Second  
 Symposium on Nonlinear Estimation Theory and Its Applications,  
 San Diego, California, Sept. 1971.
5. "Synthesis and Realization of Nonlinear Filters," Ph.D. Dissertation,  
 University of Southern California, January, 1972.

KENNETH D. SENNE

B.S. Electrical Engineering, 1964, Stanford University, California.

M.S. Electrical Engineering, 1966, Stanford University, California.

Ph.D. Electrical Engineering, 1968, Stanford University, California.

1964-66: Teaching Assistant, Stanford University, California

1965 Systems Engineer, I.B.M., Los Gatos, California  
Summer :

1966-68: Research Assistant, Stanford University, California

1968- : Research Associate, Frank J. Seiler Research  
Laboratory, Air Force Systems Command, U.S. Air  
Force Academy, Colorado

1971- : Assistant Professor of Astronautics and Computer  
Science, U. S. Air Force Academy, Colorado

Publications:

1. "Adaptive Linear Discrete-Time Estimation," Stanford University, Ph.D. dissertation, Dept. of Electrical Engineering, June 1968. Revised and edited as Stanford University Center for Systems Research, Tech. Report No. 6778-5, June 1968.
2. "Adaptive Discrete-Time Estimation," with B. Goode, Proceedings of the Purdue Centennial Year Symposium on Information Processing, Vol. 2, pp. 569-578, April 1969.
3. "Passive Receiver Design via Non-Linear Filtering," with R. Bucy and R. Geesey, Proceedings of the Third Hawaii International Conference on System Sciences, Part I, pp. 477-480, Jan. 1970.
4. "Digital Realization of Optimal Discrete-Time Nonlinear Estimators," with R. Bucy, Proceedings of the Fourth Annual Princeton Conference on Information Sciences and Systems, Princeton University, pp. 280-284, March 1970.
5. "New Results in Adaptive Estimation Theory," FJSRL Technical Report No. SRL 70-0013, April 1970.

KENNETH D. SENNE (Cont'd)

6. "Adaptive Estimation in a Nonstationary Environment," with B. Goode, presented at the 1970 International Symposium on Information Theory, Noordwijk, the Netherlands, June 1970.
7. "Synthesis of Optimum Discrete-Time Nonlinear Estimators," with R. Bucy, presented at the 1970 International Symposium on Information Theory, Noordwijk, the Netherlands, June 1970.
8. "Digital Synthesis of Nonlinear Filters," with R. Bucy, FJSRL Technical Report No. SRL 70-0010, July 1970.
9. "An Exact Solution to an Adaptive Linear Estimation Problem," FJSRL Technical Report No. SRL-TR-70-0014, Sept. 1970.
10. "Realization of Optimum Discrete-Time Nonlinear Estimators," with R. Bucy, Proceedings of the Symposium on Nonlinear Estimation Theory and Its Applications, pp. 6-17, Sept. 1970.
11. "A Unified Theory of Adaptive Estimation," Proceedings of the Fourth Hawaii International Conference on System Sciences, pp. 293-294, Jan. 1971.
12. "A Passive Ranging Experiment Using Nonlinear Filtering," with R. Bucy, Air Force Weapons Laboratories Computer Films No. 71-0330-02, March 1971.
13. "Digital Synthesis of Nonlinear Filters," with R. Bucy, Automatica, Vol. 7, pp. 287-299, May 1971.
14. "Computer Experiments with Nonlinear Filters," Proceedings of the Second Symposium on Nonlinear Estimation Theory and Its Applications, pp. 314-324, Sept. 1971.
15. "Digital Implementation of Bayes Law for Nonlinear Estimation," presented at the 1971 IEEE Conference on Decision and Control, 15-17 Dec 1971.
16. "An Optimal Phase Demodulator," with R. Bucy, Air Force Weapons Laboratories Computer Films No. 72-0130-01, January 1972.
17. "Optimal Phase Demodulation via Discrete Nonlinear Filtering," with R.S. Bucy and C. Hecht, Air Force Weapons Laboratories Computer Films No. 72-0401-01, April 1972.
18. "Bayes Law Implementation: Optimal Discrete-Time Phase Estimation," presented at the 1972 SWIEEEO, Dallas, Texas, 19-21 April 1972.

END  
FILMED  
9-23-72  
NTIS